Nikolaj Theodor Birkmose Thams

Causality and Distribution Shift

PHD THESIS

THIS THESIS HAS BEEN SUBMITTED TO THE PHD SCHOOL OF
THE FACULTY OF SCIENCE, UNIVERSITY OF COPENHAGEN

Department of Mathematical Sciences
University of Copenhagen

August 2022

Nikolaj Theodor Birkmose Thams
thams@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Causality and Distribution Shift
Supervisor:	Professor Jonas Peters University of Copenhagen
Assessment Committee:	Professor Susanne Ditlevsen (chair) University of Copenhagen
	Professor Thomas Richardson University of Washington
	Professor Rajen Shah University of Cambridge
Date of Submission:	August 31, 2022
Date of Defense:	November 4, 2022
ISBN:	978-87-7125-061-9

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen. It was supported by the Villum Foundation (research grant 18968).

Preface

The work contained in this PhD thesis is the result of three wonderful years of PhD studies under the supervision of Professor Jonas Peters at the Department of Mathematical Sciences, University of Copenhagen.

I am grateful to Jonas for his enormous support over the last three years, for his happy and playful approach to doing research, and for all the attention, time and effort he dedicates to his students.

I thank my colleagues and now friends at the University of Copenhagen. Coming to the office has been a daily joy, and I have found great inspiration in our discussions about statistics and beyond. I especially want to thank Niels Richard, for the mentorship that inspired me to do a PhD. I am thankful to David, Mike, Stefan and the rest of the Clinical ML lab at MIT for being so welcoming during my stay in Boston.

I am grateful to my family and friends for their encouragement during my studies, in particular to Frederik, Peter and Lau, with whom I have shared the ups and downs of doing a PhD.

Finally, I am forever grateful to Amalie, for the interest, patience, love and support she continuously shows me.

Nikolaj Theodor Birkmose Thams
August, 2022

Abstract

This PhD thesis contains a number of contributions on drawing causal inference from observational data, with a particular focus on shifts in distribution. These contributions fall within three categories: 1) Testing hypotheses in shifted distributions, 2) learning predictive models that are robust to distribution shift and 3) inferring causal structure and causal effects using exogenous variables.

First, we present a general framework which formalizes statistical hypothesis testing under distribution shifts. We propose methods and prove theoretical results for conducting such tests. We describe a number of different applications of testing under distribution shifts, which includes policy learning and conditional independence testing.

Second, we outline ways of using causal methodology to learn predictive models that are robust to shifts in distribution. We propose an algorithm for learning invariant policies in bandit problems, and we show that, if certain assumptions are satisfied, this allows for worst-case optimal prediction in unseen environments. In a regression setting, we propose an estimator for learning linear predictors that are worst-case optimal over a class of mean-shifts in an unobserved confounder, assuming that we observe proxies of this confounder. We also propose a framework for specifying plausible parametric shifts in distribution, and develop theory for finding the shift that has the worst-case impact on the performance of a predictive model.

Finally, we provide methods for inferring causal structure and causal effects from heterogeneous observational data. We propose a procedure for identifying causal ancestors of a given target variable by using 'minimal invariance' of sets of predictors across multiple exogenous environments (or distribution shifts). We develop instrumental variable methodology for inferring causal effects in linear time series data, where we highlight that past states are helpful for obtaining 'more exogeneity' but also that past states confound the instrument and outcome, and needs to be adjusted for.

Sammenfatning

Denne Ph.D.-afhandling indeholder en række bidrag vedrørende kausal inferens draget fra observationelle data, med et særligt fokus på skift i fordeling. Bidragene falder indenfor tre kategorier: 1) At teste hypoteser i skiftede fordelinger, 2) at lære prædiktive modeller der er robuste overfor skift i fordelinger og 3) at inferere kausale strukturer og kausale effekter ved brug af exogene variable.

Først præsenterer vi en generel formalisering af statistiske hypotesetest under skift i fordeling. Vi foreslår metoder og beviser teoretiske resultater, der muliggør sådanne test. Vi beskriver en række forskellige anvendelser af test under skift i fordeling, hvilket inkluderer evaluering af strategier i beslutningsproblemer og test af betinget uafhængighed.

Dernæst beskriver vi hvordan kausale metoder kan bruges til at lære prædiktive modeller, som er robuste overfor skift i fordeling. Vi foreslår en algoritme, der kan lære invariante strategier i 'bandit' problemer, og viser at denne, under visse antagelser, muliggør worst-case optimale prædiktioner i nye miljøer. Vi foreslår en regressionsestimator til at lære lineære modeller som er worst-case optimale over en klasse af middelværdiskift i en uobserveret confounder, under antagelse af, at vi observerer en proxy for denne confounder. Og vi beskriver et system til at specificere plausible parametriske skift i fordeling, og foreslår en metode til at identificere hvilket skift, der har den værste indvirkning på en prædiktiv models præstation.

Endelig fremlægger vi metoder til at inferere kausal struktur og kausale effekter fra heterogene observationelle data. Vi foreslår en procedure til at inferere kausale forfædre til en responsvariabel ved at bruge 'minimal invarians' på tværs af adskillige miljøer (eller skift i fordeling). Og vi udvikler metoder der infererer kausale effekter i lineære tidsrækker ved brug af instrumental variable. Vi fremhæver, at tidligere tilstande kan bruges til at øge 'mængden af exogenitet', men understreger også vigtigheden af at justere for tidligere tilstande som confoundere.

Contributions and Structure

This thesis contains an introduction in Chapter 1, which briefly reviews causal inference. The introduction is not intended to be a representative review of the literature, but rather a highly selective summary of methods which the later chapters build upon. It is followed by 3 chapters, each of which contains a small motivation for the problems studied as well as one or more papers. For reference within this thesis, we give each paper an acronym, for example **[ShiftTest]**. All theorems, etc., are numbered relative to the paper they appear in.

Chapter 2 (Shifts in Distribution: Testing) discusses hypothesis testing in an unobserved target distribution P using data observed from a different distribution Q. The chapter contains the following paper:

[ShiftTest] [Thams et al., 2021]. N. Thams, S. Saengkyongam, N. Pfister, and J. Peters. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

Paper status: Revision under review at JRSS-B.

Chapter 3 (Shifts in Distribution: Prediction) is concerned with prediction under distribution shift, where a model is trained in one distribution Q, but will also be applied in one or more different distributions. The chapter contains the following three papers:

[ShiftEval] [Thams et al., 2022a]. N. Thams, M. Oberst, and D. Sontag. Evaluating robustness to dataset shift via parametric robustness sets. In *Neural Information Processing Systems (NeurIPS)*, 2022a. NT and MO contributed equally, order determined by coin flip.

Paper status: Under review at NeurIPS 2022.

[ProxyAR] [Oberst et al., 2021]. M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

[InvPolicy] [Saengkyongam et al., 2021]. S. Saengkyongam, N. Thams, J. Peters, and N. Pfister. Invariant policy learning: A causal perspective. *arXiv preprint arXiv:2106.00808*, 2021.

Paper status: Revision in progress for resubmission at IEEE TPAMI.

Chapter 4 (Shifts in Distribution: Causal Inference) considers estimation of causal structure and causal effects from observational data with exogenous environments or instruments. The chapter contains the following two papers:

[TimeIV] [Thams et al., 2022b]. N. Thams, R. Søndergaard, S. Weichwald, and J. Peters. Identifying causal effects using instrumental time series: Nuisance IV and correcting for the past. *arXiv preprint arXiv:2203.06056*, 2022b.

Paper status: Revision in progress for resubmission at JMLR.

[AncSearch] [Mogensen et al., 2022]. P. Mogensen, N. Thams, and J. Peters. Invariant ancestry search. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15832–15857. PMLR, 2022.

During my PhD, I also worked on the following two papers, which however are not included in this thesis.

1. N. Thams and N. R. Hansen. Local independence testing for point processes. *arXiv preprint arXiv:2110.12709*, 2021.

2. S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pages 27–36. PMLR, 2020

Contents

Preface iii

Abstract iv

Contributions and Structure vii

1. Introduction 1
- 1.1. Structural Causal Models . . . 2
- 1.2. Causal Models when Graphs are Known: Effect Estimation and Instrumental Variables . . . 5
- 1.3. Causal Models when Graphs are Unknown: Invariance and Exogeneity . . 9
- 1.4. Causal Models in Distribution Shift . . . 10

2. Shifts in Distributions: Testing 13

Statistical Testing under Distributional Shifts 15
- 1. Introduction . . . 15
- 2. Statistical Testing under Distributional Shifts . . . 18
- 3. Example Applications of Testing under Distributional Dhifts . . . 21
- 4. Testing by Resampling . . . 29
- 5. Experiments . . . 41
- 6. Conclusion and Future Work . . . 56

3. Shifts in Distributions: Prediction 59

Evaluating Robustness to Dataset Shift via Parametric Robustness Sets 61
- 1. Introduction . . . 61
- 2. Defining Parametric Robustness Sets . . . 64
- 3. Evaluation of the Worst-Case Loss . . . 66
- 4. Experiments . . . 70
- 5. Discussion . . . 73

Regularizing towards Causal Invariance: Linear Models with Proxies 75
- 1. Introduction . . . 75
- 2. Preliminaries . . . 77
- 3. Distributional Robustness to Bounded Shifts . . . 79
- 4. Targeted Anchor Regression: Incorporating Additional Shift Information . 83
- 5. Synthetic Experiments . . . 86

6. Real-Data Experiment: Pollution . . . 88
7. Discussion and Related Work . . . 90

Invariant Policy Learning: A Causal Perspective **93**
1. Introduction . . . 93
2. A Causal Framework for Multi-environment Contextual Bandits . . . 97
3. Invariant Policies for Distributional Robustness . . . 104
4. Learning an Optimal Invariant Policy . . . 107
5. Simulation Experiments . . . 114
6. Warfarin Dosing Case Study . . . 116
7. Conclusion . . . 120

4. Shifts in Distributions: Causal Inference **123**

Invariant Ancestry Search **125**
1. Introduction . . . 125
2. Preliminaries . . . 127
3. Minimal Invariance and Ancestry . . . 128
4. Oracle Algorithms . . . 129
5. Invariant Ancestry Search . . . 131
6. Experiments . . . 134
7. Extensions . . . 140
8. Conclusion and Future Work . . . 140

Identifying Causal Effects using Instrumental Time Series: Nuisance IV and Correcting for the Past **143**
1. Introduction . . . 143
2. Causal Time Series Models with Confounding . . . 147
3. Nuisance Effects in Instrumental Variable Regression . . . 150
4. Instrumental Time Series Regression . . . 153
5. Simulation Experiments . . . 161
6. Conclusion and Future Work . . . 165

A. Appendix to Statistical Testing under Distributional Shifts **167**

B. Appendix to Evaluating Robustness to Dataset Shift via Parametric Robustness Sets **205**

C. Appendix to Regularizing towards Causal Invariance: Linear Models with Proxies **245**

D. Appendix to Invariant Policy Learning: A Causal Perspective **275**

E. Appendix to Invariant Ancestry Search **299**

F. Appendix to Identifying Causal Effects using Instrumental Time Series: Nuisance IV and Correcting for the Past **321**

Bibliography **349**

1. Introduction

In this PhD thesis, we develop theory for drawing causal inference from observational data, and we analyse several aspects of distribution shift and their relations to causality. In broad terms, causal inference aims to learn a mechanistic understanding of how data is generated; such understanding exceeds a simple description of the observed joint distribution of the data, in that it also explains why the data behave as they do.

For example, and perhaps most famously, causal inference has been applied to remove confounding effects, which we define as statistical dependence between two variables due to a common cause. A frequently studied example of such confounding is the association between the dose of a drug given to a patient, and the outcome observed in the patient; among others, this is confounded by the severity of the disease. If we plainly considered the observed association between the dose and the outcome, we may underestimate the effect of the drug, because the patients that receive the largest doses may also be the patients that are most sick and thus, in spite of the increased dose, suffer from worse outcomes. Causal inference provides methodology for adjusting for confounding, to obtain estimates of the 'causal effect' of the drug, which is to be understood as the effect of the drug, if confounding was not present.

Yet, causal inference applies to other tasks than estimating causal effects through removal of confounding. In this thesis, we explore ways of using causal methods in the presence of distribution shift. The term 'distribution shift' broadly refers to a setting, where we do not just train models and predict outcomes in a single distribution Q, but instead we may observe training data from several different distributions, or we may observe training data from one distribution but want to predict in a different, unobserved distribution. For example, in [**InvPolicy**], we observe data from a number of different distributions (or 'environments'), $Q_1, \ldots, Q_d$, and aim to learn a model that performs well, not only in $Q_1, \ldots, Q_d$, but also in a number of unobserved environments $P_1, \ldots, P_m$.

A motivation for considering distribution shift in the context of causality is to use 'causality's mechanistic understanding' to learn models that generalize better to test distributions. This motivation relies on the assumption that causal dependencies between a predictor and an outcome is stable, and likely to be the same in both training and test environments, whereas confounding and spurious correlations are less stable. As a result, a model which relies too much on non-causal features, may fail in deployment. While this is only an assumption (until we encode it into a concrete data generating mechanism, in which case it may become an over-simplification), it is appealing to believe that causal effects are, if not perfectly stable across time and environments, then at least more stable than confounding. For example, while the causal effect of a drug may be roughly the same in different populations (e.g. from different health care systems),

confounding factors, such as treatment allocation, may differ substantially between those populations.

The remainder of this introduction contains a selection of existing results that this thesis builds upon, and it is structured as follows. In Section 1.1, we discuss structural causal models (SCMs). Throughout the thesis, SCMs will be the statistical framework that we assume generates data, and in some cases also the target of inference that we aim to partially infer from data. In Section 1.2, we discuss causal methodology when a causal graph, that is a hierarchy of cause and effect, is known. Of particular interest to us is instrumental variables (IV) regression as a method of learning causal effects despite unobserved confounding. In Section 1.3, we discuss causal methodology when no causal graph is known, and in particular we discuss invariant causal prediction (ICP) and how to use causal exogeneity to learn causal structure. Finally, in Section 1.4, we discuss shifts in distribution, how they relate to causal interventions and how to use importance sampling to estimate means in shifted distributions.

1.1. Structural Causal Models

To describe causal relationships, the papers included in this thesis all utilize the framework of structural causal models or SCMs [Pearl, 2009, Peters et al., 2017], which we now review. An SCM $\mathfrak{C}$ over variables $X_1, \dots, X_d$ is given by the assignments

$$X_j := f_j(\mathrm{PA}_j, \varepsilon_j) \tag{1}$$

where for each j, f_j is some function, $\mathrm{PA}_j \subseteq \{X_1, \dots, X_d\}$ is a collection of variables (which we call the *parents* of j), and $\varepsilon_1, \dots, \varepsilon_d$ is a collection of jointly independent noise variables each with distribution $\mathbb{P}_j$.

The assignments in (1) induces a joint distribution $P^{\mathfrak{C}}$ over $X_1, \dots, X_d$, which we refer to as the *observational distribution*. However, assuming that variables $X_1, \dots, X_d$ are generated according to an SCM $\mathfrak{C}$ is stronger than assuming that data follows a joint distribution $P^{\mathfrak{C}}$, since the SCM also describes the mechanistic relationships generating this data; for example, the SCM also induces interventional distributions.

Given an SCM $\mathfrak{C}$, we can form an intervened SCM, $\tilde{\mathfrak{C}} := \mathfrak{C}^{\mathrm{do}(X_j := \tilde{f}_j(\widetilde{\mathrm{PA}}_j, \tilde{\varepsilon}_j))}$, by changing the assignment of X_j in (1) into $X_j := \tilde{f}_j(\widetilde{\mathrm{PA}}_j, \tilde{\varepsilon}_j)$ for some function $\tilde{f}_j$, parent set $\widetilde{\mathrm{PA}}_j$ and noise variable $\tilde{\varepsilon}_j$, and where we let the assignments of all other variables $X_i, i \neq j$, in (1) remain unchanged. We call the intervention a *hard intervention* if $\widetilde{\mathrm{PA}}_j = \emptyset$ and $\tilde{f}_j(\widetilde{\mathrm{PA}}_j, \tilde{\varepsilon}_j) \overset{a.s.}{=} x$ and simply write $\mathrm{do}(X_j := x)$. The intervened SCM, $\tilde{\mathfrak{C}}$, induces a distribution $P^{\tilde{\mathfrak{C}}}$ over $X_1, \dots, X_d$, which we call the *interventional distribution*.

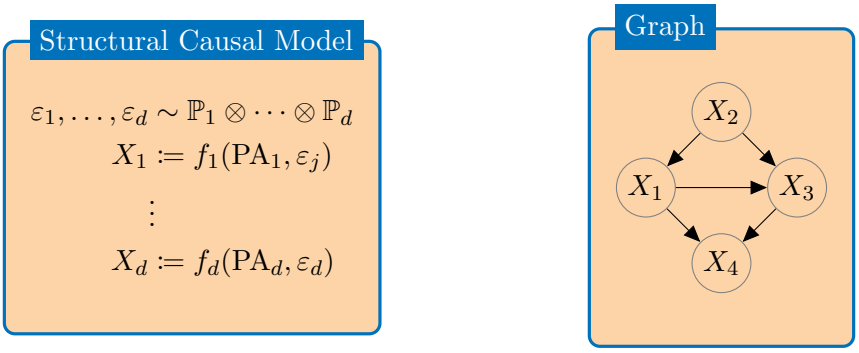


Figure 1.: (Left) An SCM is a set independent error variables ε_j, parent sets $\mathrm{PA}_j \subseteq \{X_1, \dots, X_d\}$ that specify direct generative dependencies and functional assignments f_j. (Right) An SCM induces a graph over the variables $X_1, \dots, X_d$ (sometimes written $1, \dots, d$), where an edge $X_i \to X_j$ indicates that $X_i \in \mathrm{PA}_j$.

Example 1. *Consider the following SCM:*

$$\begin{aligned} X_2 &\coloneqq \varepsilon_2, \\ X_1 &\coloneqq \mathbb{1}_{\{\sigma(X_2) > \varepsilon_1\}}, \\ X_3 &\coloneqq \theta \cdot X_1 + X_2 + \varepsilon_3, \\ X_4 &\coloneqq X_3 + X_1 + \varepsilon_4, \end{aligned}$$

where $\varepsilon_1 \sim \mathit{Unif}(0,1)$, $\varepsilon_2, \varepsilon_3, \varepsilon_4 \sim \mathcal{N}(0,1)$, $\varepsilon_1, \dots, \varepsilon_4$ *are jointly independent,* σ *is the logistic sigmoid function and* $\theta \in \mathbb{R}$. *Here* $X_1 \in \{0,1\}$ *is a binary variable whose probability of being* 1 *is* $\sigma(X_2)$. *In Fig. 1 (right) we draw a graph corresponding to these assignments.*

When conditioning on $X_1 = 1$, *we have that* $\mathbb{E}_P[X_2 | X_1 = 1] = 2\mathbb{E}_P[X_2 \sigma(X_2)] \approx 0.4$,[1] *so in the observational distribution,*

$$\mathbb{E}_P[X_3 | X_1 = 1] = \theta \mathbb{E}_P[X_1 | X_1 = 1] + \mathbb{E}_P[X_2 | X_1 = 1] + \mathbb{E}_P[\varepsilon_3 | X_1 = 1] \approx \theta + 0.4,$$

where we use that $\mathbb{E}_P[\varepsilon_3 | X_1 = 1] = \mathbb{E}_P[\varepsilon_3] = 0$, *since* $\varepsilon_3 \perp\!\!\!\perp X_1$. *Similarly* $\mathbb{E}_P[X_3 | X_1 = 0] \approx -0.4$, *and despite that* X_1 *only enters the structural equation for* X_3 *with a coefficient of* θ, *the difference of conditioning on* $X_1 = 1$ *and* $X_1 = 0$ *is approximately* $\theta + 0.8$, *because the value of* X_1 *carries additional confounded information, which change the expected value of* X_3.

We can also consider the interventional distribution $\tilde{P} = P^{\mathrm{do}(X_1 := 1)}$. *In this distribution the confounding between* X_1 *and* X_3 *is removed because* X_1 *and* X_2 *are now independent, and*

$$\mathbb{E}[X_3 | \,\mathrm{do}(X_1 \coloneqq 1)] = \theta \mathbb{E}_{\tilde{P}}[X_1] + \mathbb{E}_{\tilde{P}}[X_2] + \mathbb{E}_{\tilde{P}}[\varepsilon_3] = \theta,$$

[1] This follows from Bayes rule $p(x_2 | X_1 = 1) = p(X_1 = 1 | x_2) \frac{p(x_2)}{p(X_1 = 1)} = 2\sigma(x_2) p(x_2)$, where p is the density (with respect to the product of a Lebesgue and a discrete measure).

where we use the notation $\mathbb{E}[X_3|\,\mathrm{do}(X_1 := 1)] := \mathbb{E}_{\tilde{P}}[X_3]$ *and that* $\mathbb{E}_{\tilde{P}}[X_2] + \mathbb{E}_{\tilde{P}}[\varepsilon_3] = \mathbb{E}_P[X_2] + \mathbb{E}_P[\varepsilon_3] = 0$ *because the intervention neither changes* X_2 *nor* ε_3*. Similarly,* $\mathbb{E}[X_3|\,\mathrm{do}(X_1 = 0)] = 0$*, and the difference of intervening* $\mathrm{do}(X_1 := 1)$ *and* $\mathrm{do}(X_1 := 0)$*, is* θ*, which we call the* causal effect *of* X_1 *on* X_2*. This example highlights that conditioning differs from intervening, for example in that the conditioning includes correlation due to confounders, while the intervention breaks this confounding.*

An intervention setup like this can, for example, be used to describe a patient's response, X_3, to a treatment, X_1, when a confounder X_2 affects both the treatment and the outcome. Here we may consider the intervention $\mathrm{do}(X_1 := 1)$ to study what will happen if everyone is given the drug, or the intervention $\mathrm{do}(X_1 := \mathrm{Bern}(0.5))$, which corresponds to the distribution one will get in a randomized control trial (RCT).

Graph Terminology

We visualize the causal hierarchy of an SCM by drawing a directed graph $\mathcal{G} = (V, E)$ over nodes $V = \{1, \dots, d\}$ (or interchangeably over nodes $V = \{X_1, \dots, X_d\}$) and edges E, where an edge $(i, j) \in E$ if $X_i \in \mathrm{PA}_j$ (we also denote this $i \to j$). It is common to assume that the graph of an SCM is a directed acyclic graph (DAG), which ensures that a causal ordering from 'first' to 'last' exists although we can define causal models without this assumption, see Bongers et al. [2021]. We now review some graph terminology and some basic concepts of graphical models for a DAG $\mathcal{G}$.

If $i \to j$, we say that i is a *parent* of j and that j is a *child* of i. A *path* π between v_0 and v_m is a sequence of vertices and edges $\pi = (v_0, e_1, v_1, \dots, e_m, v_m)$ where either $e_k = (v_{k-1}, v_k)$ or $e_k = (v_k, v_{k-1})$; we sometimes simply write $\pi = (e_1, \dots, e_m)$. A *directed path* from v_0 to v_m is a path $\pi = (v_0, e_1, \dots, v_m)$ where $e_k = (v_{k-1}, v_k)$, that is all edges have the same orientation. A node $i \in V$ is an *ancestor* of $j \in V$ if a directed path from i to j exists, and similarly j is then a *descendant* of i. We denote the parents, children, ancestors and descendants of j by $\mathrm{PA}_j, \mathrm{CH}_j, \mathrm{AN}_j$ and DE_j respectively. For a set $C \subseteq V$, we write $\mathrm{AN}_C = \cup_{j \in C} \mathrm{AN}_j$ (and similarly for $\mathrm{PA}_C, \mathrm{CH}_C$ and DE_C).

If a path $\pi = (v_0, e_1, \dots, v_m)$ contains a segment $v_{k-1} \overset{e_k}{\to} v_k \overset{e_{k+1}}{\leftarrow} v_{k+1}$, we say that v_k is a *collider* on π, and otherwise we say that v_k is a non-collider. A path π from i to j is *d-connected* (or open) given a set $C \subseteq V \setminus \{i, j\}$ if every non-collider on π is not in C and every collider on π is in $C \cup \mathrm{AN}_C$. Otherwise we say that π is *blocked* by C. If all paths between i and j are blocked by a set $C \subseteq V \setminus \{i, j\}$, we say that i and j are *d-separated* by C, in which case we write $i \perp j | C$.

For example, in Fig. 1, the path $X_1 \to X_4 \leftarrow X_3$ is d-connected given X_4, but is blocked by $\emptyset$. X_1 and X_3 are not d-separated given any conditioning set, since for any set C, the path $X_1 \to X_3$ is open given C. On the contrary, X_2 and X_4 are d-separated by $\{X_1, X_3\}$, since all paths between X_2 and X_4 use either X_1 or X_3 as a non-collider.

The Global Markov Property and Causal Factorizations

Given a joint distribution P over variables $X_1, \dots, X_d$, we write $X_i \perp\!\!\!\perp X_j \mid X_C$ if X_i and X_j are conditionally independent given X_C. A joint distribution satisfy the *global Markov*

property with respect to a graph $\mathcal{G}$ if for all $i \neq j$ and $C \subseteq V \setminus \{i, j\}$, $X_i \perp X_j | X_C$ implies that $X_i \perp\!\!\!\perp X_j | X_C$. That is, d-separations in the graph imply conditional independences in the distribution. The reverse property is called *faithfulness*: P is faithful with respect to $\mathcal{G}$ if for all $i \neq j$ and $C \subseteq V \setminus \{i, j\}$, $X_i \perp\!\!\!\perp X_j | X_C$ implies that $X_i \perp X_j | X_C$.

For example, in the graph in Fig. 1 (right), $X_2 \perp X_4 | \{X_1, X_3\}$, and if the global Markov property is satisfied for the distribution then we can conclude conditional independence $X_2 \perp\!\!\!\perp X_4 | \{X_1, X_3\}$. There exist several different conditions which ensure the global Markov property, such as when data is sampled from an SCM where the observed distribution has density with respect to a product measure [Peters et al., 2017], and all the papers in this thesis uses the global Markov property, implicitly or explicitly. On the contrary, faithfulness is in many cases not required to make causal statements, though it often is required when learning causal structures – for example, we do assume faithfulness in **[AncSearch]** to learn ancestral sets.

A consequence of the global Markov property is that the joint distribution P over variables $X_1, \ldots, X_d$ that are generated by an SCM $\mathfrak{C}$ factorizes according to the parent sets. If p_P is the density of P, then

$$p_P(x_1, \ldots, x_d) = \prod_{j=1}^{d} p_P(x_j | x_{\mathrm{PA}_j}), \tag{2}$$

which we call the causal factorization [Pearl, 2009]. When intervening on a variable X_j, $\tilde{\mathfrak{C}} = \mathfrak{C}^{\mathrm{do}(X_j := \tilde{f}_j(\widetilde{\mathrm{PA}}_j, \tilde{\varepsilon}_j))}$, all factors in (2) remain the same, except the one relating to X_j

$$p_{\tilde{P}}(x_1, \ldots, x_d) = \prod_{i=1}^{d} p_{\tilde{P}}(x_i | x_{\mathrm{PA}_i}) = \left(\prod_{\substack{i=1 \\ i \neq j}}^{d} p_P(x_i | x_{\mathrm{PA}_i}) \right) p_{\tilde{P}}(x_j | x_{\widetilde{\mathrm{PA}}_j}), \tag{3}$$

where $p_{\tilde{P}}$ is the density of the intervened distribution $\tilde{P}$. There are several other ways we can factorize the joint distribution, for example $p_P(x_1, \ldots, x_d) = \prod_{i=1}^{d} p_P(x_i | \{x_j\}_{j<i})$, but in such a 'non-causal' factorization, an intervention in most cases will change several, if not all, factors in the product. In Section 1.4 we discuss how this factorization is useful in distribution shifts when various distributions are generated from the same SCM but with different interventions.

1.2. Causal Models when Graphs are Known: Effect Estimation and Instrumental Variables

The graph of an SCM contains less information about a distribution than the full SCM, for example because the graph does not specify the functional connections f_j. Yet, we can make different causal statements and use different methodology depending on whether or not we assume that the graph is known. For example, if the graph is known and all variables are observed and we additionally assume additive mean-zero noise, we

can in principle consistently recover the functions $f_1, \dots, f_d$, by non-linearly regressing X_j onto X_{PA_j} for each j (and assuming that our model class is rich enough to capture f_j); effectively we have thereby learned the entire SCM.

This, however, does not mean that causal inference with known graphs is trivial, and there are many questions to be studied, such as how to efficiently estimate causal effects or how to do so in the presence of unobserved confounders. Many of these questions can be addressed in the context of do-calculus [Pearl, 2009], which is a set of rules for how to convert statements about interventional distributions into statements about observational distributions. One classic application of do-calculus is computing intervention effects using covariate adjustment.

Covariate Adjustment

Consider an SCM $\mathfrak{C}$ over variables X, Y and C with the graph in Fig. 2 (left). Suppose we want to compute the mean of Y in an intervention $\tilde{\mathfrak{C}} = \mathfrak{C}^{\mathrm{do}(X:=x_0)}$ without observing any data from $\tilde{P}$. We can use *covariate adjustment*, by adjusting for C [Robins, 1986, Pearl, 2009]. If we assume that all variables are discrete, we have

$$\begin{aligned}
\mathbb{E}_{\tilde{P}}[Y] &= \sum_{y,c} y \cdot \mathbb{P}_{\tilde{P}}(Y = y, C = c, X = x_0) \\
&= \sum_{y,c} y \cdot \mathbb{P}_{\tilde{P}}(Y = y | C = c, X = x_0) \cdot \mathbb{P}_{\tilde{P}}(C = c, X = x_0) \\
&= \sum_{y,c} y \cdot \mathbb{P}_{P}(Y = y | C = c, X = x_0) \cdot \mathbb{P}_{P}(C = c)
\end{aligned}$$

In the last equation, we use that the conditional distribution of a non-intervened variable given its parents does not change between P and $\tilde{P}$. This calculation shows that we can sometimes derive quantities about the interventional distribution $\tilde{P}$ from the observational distribution P alone.

Many other works in the literature have studied similar questions of how to reason about intervention effects, without observing any data from $\tilde{P}$ [e.g. Tian and Pearl, 2002, Pearl, 2009]. A related question is that of transportability, introduced by Pearl and Bareinboim [2011]. Given a target distribution P that differs in some (known) way from a training distribution Q, they infer effects of interventions in P from the observational and the interventional distribution of Q.

Instrumental Variables

Sometimes, adjusting for covariates may not be possible, for example if C is not observed. Instead, in some cases, we can rely on instrumental variable (IV) approaches to estimate causal effects.

In Example 1, we considered a causal effect defined as $\mathbb{E}[X_3|do(X_1 := 1)] - \mathbb{E}[X_3|do(X_1 := 0)]$. If we consider continuous variables, we may instead define the causal effect of X on Y as $\frac{\partial}{\partial x}\mathbb{E}^{\mathrm{do}(X:=x)}[Y]$. While this is in general a function of x, it becomes constant if the

Figure 2.: (Left) Graph where we can use covariate adjustment to reason about the causal effect interventions on X have on Y. (Right) Graph where we can use instrumental variables to reason about the causal effect interventions on X have on Y. The dashed circle around C indicates that it is unobserved.

assignment of Y is additive and linear in X, because in that case $\mathbb{E}^{\text{do}(X:=x)}[Y] = \beta x + K$ for a constant K, and we call β the linear causal effect of X on Y. Instrumental variables, which we will discuss now, can be used for estimating such linear causal effects (though non-linear functions can also be estimated [Christiansen et al., 2021]).

When inferring the causal effect of X on Y, an instrumental variable, Z, is a variable which satisfies the following graphical conditions [Pearl, 2009].

1. Z is d-separated from Y in the graph $\mathcal{G}_{X \not\to Y}$ where we remove any edge $X \to Y$,
2. Z is not d-separated from X and
3. Z is d-separated from C.

For example in Fig. 2 (right), Z is an instrumental variable for estimating the causal effect of X on Y: Z and X are not d-separated due to the edge $Z \to X$, Z and C are d-separated due to the colliders $C \to X \leftarrow Z$ and $C \to Y \leftarrow X$, and when removing the edge $X \to Y$, the only path from Z to Y contains the blocked collider $Z \to X \leftarrow C$, meaning that Z and Y are d-separated after removing $X \to Y$.

Under some additional assumptions, IVs can ensure identifications of causal effects. For example, assume that $Y = \beta X + g(C, \varepsilon_Y)$, that is Y is linear in X and the contribution from (C, ε_Y) is additive. It follows from IV assumption 3. and the global Markov property that $Y - \beta X = g(C, \varepsilon_Y)$ is independent of Z. We can use this independence to identify the causal effect β, by finding vectors b which make $Y - bX$ independent of Z and we refer to such b as being *invariant*. It is common to use uncorrelatedness as a proxy for independence, and solve the resulting estimating equations [Hall, 2005]. In order to get a unique solution b that make $Y - bX$ uncorrelated to Z, the dimension of the instrument, Z, needs to be sufficiently high, typically higher than the dimension of X, which is known as an 'over-identified' setting.

Example 2. *A famous application of IV regression is that of Angrist and Krueger [1991], who use instrumental variables to estimate the causal effect of the number of years of schooling (X) on earnings (Y) in adulthood. The variables X and Y are confounded by numerous unobserved factors (C), including for example how resourceful the parents are. This means that simply regressing Y onto X and adjusting for observed confounders, will likely not result in a good estimate of the causal effect of forcing every school student to spend another year in school.*

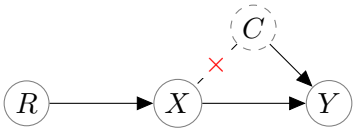


Figure 3.: Graphical illustration of the dependencies in an idealized randomized control trial, where a treatment X is made independent of confounders C through a randiomization R. Though this is not a formal graphical model, we can see similarities to the IV graph in Fig. 2.

Instead, Angrist and Krueger [1991] use the quarter of birth (Z) as an instrument. The variable Z correlates with X because of rules in the US school system which mean that the length of schooling before a student can drop out depends on a students birth day. Further, it is plausible that birth date is independent of the confounding factors C, and finally, one could argue that Z does not influence Y in other ways than through X (though this assumption is untestable).

Thus the assumptions for being a valid instrument are satisfied; Angrist and Krueger [1991] estimate the benefit of an additional year of schooling and find that this has a positive impact on future earnings.

Example 2 shows that we can think of IV as a pseudo-randomization: No randomization is conducted by a practicioner, but an external source of randomness (here the quarter of birth) still impacts the allocation of treatment; while we do not directly observe unconfounded treatments, we can 'deconfound' the treatment by using the correlation to the instrument. In Fig. 3, we visualize an idealized randomized control trial, where the randomization R breaks the dependence on a confounder C.[2] This structure is similar to the IV graph in Fig. 2 (right), with the difference that in the idealized RCT, the edge from C to X is fully broken, whereas in the IV graph this edge is present, but can be filtered out using IV methods. It is helpful (though not perfectly accurate) to think about this as that the more variance in X is explained by the instrument, the easier it is to break the confounding link, and that the idealized RCT represents the extreme of a perfectly correlated instrument.

If the instrument Z is not sufficiently high-dimensional to identify a unique parameter vector, a subspace of solutions $M = \{b \in \mathbb{R}^{d_X} \mid \mathrm{cov}(Y - bX, Z) = 0\}$ exists, which is known as the 'under-identified' setting. To chose a single regression model from M, we can search among all vectors b that (approximately) satisfy invariance, and select the one which yields the smallest mean squared prediction error. Rothenhäusler et al. [2021] do this by minimizing a convex combination of an invariance loss and a predictive error, whereas Jakobsen and Peters [2021] select the most predictive model among all those that are not significantly non-invariant.

[2] We only draw this graph to highlight the similarity to the IV graph, but we should be careful in interpreting this as a graphical model arising from an SCM; for example, if R and X are both binary, they may essentially be the same random variable, which can be problematic in a graphical model; we also have not defined what the crossed out edge means. One can draw more formal graphs for this setup using the framework in Richardson and Robins [2013].

1.3. Causal Models when Graphs are Unknown: Invariance and Exogeneity

In many applications, a causal graph may not be available, and instead we only observe data from an observational distribution $P^{\mathfrak{C}}$. Various methods can be used to estimate causal graphs from data; for example, constraint based methods use conditional independences to learn an estimate of a graph [Spirtes et al., 2000, Chickering, 2002]. In general we cannot recover the graph of $\mathfrak{C}$ from $P^{\mathfrak{C}}$, but only the *Markov equivalence class* of the graph. Two DAGs are Markov equivalent if they have the same skeleton (i.e. the same edges, when orientations are removed) and the same v-structures [Verma and Pearl, 1991]. Three nodes x, y, z in a graph $\mathcal{G}$ form a v-structure if the path $x \to y \leftarrow z$ is in $\mathcal{G}$ and neither $x \to z$ nor $x \leftarrow z$ is in $\mathcal{G}$. Constraint based methods do not assume any model class (though in practice, they require that data follows a model class for which powerful conditional independence tests are available), but if we are willing to make assumptions, such as non-Gaussianity [Shimizu et al., 2006] or non-linearity [Hoyer et al., 2009], we can learn not just the Markov equivalence class, but the graph of the SCM. Yet, estimating entire graphs from data remains a difficult task, and for many applications, it may be sufficient to learn parts of a graph, such as, if we only seek to learn causal parents of a target variable Y.

Invariant Causal Prediction

Instead of learning the full graph, we can in some cases use limited graphical information, such as exogeneity of a variable E (we say that E is exogenous if $\mathrm{PA}_E = \emptyset$), to learn causal structure.

Example 3. *Suppose we observe variables E, X, Y from an SCM where the only non-trivial (conditional) independence relation between them is $E \perp\!\!\!\perp Y | X$. From this we can infer that the graph of the SCM must have the skeleton $E - X - Y$, and no v-structure, but per the characterization of Markov equivalence classes above, we cannot distinguish between the three graphs*

$$E \to X \to Y$$
$$E \leftarrow X \to Y$$
$$E \leftarrow X \leftarrow Y.$$

However, if we additionally know that E is exogenous, this not only tells us the direction of the edge $E \to X$, but also that $X \to Y$.

Example 3 shows that we can leverage knowledge of exogeneity of one variable to direct edges between other variables. A structured exploitation of this fact is used by Invariant Causal Prediction (ICP) by Peters et al. [2016], who do not assume knowledge of a causal graph, but only assumes the presence of exogenous environments, E. They consider a target variable Y and predictors $X_1, \ldots, X_d$, and assume that the conditional distribution of Y given its causal parents does not change across environments; if the

environments correspond to different interventions, this assumption corresponds to assuming that Y is not the target of intervention (see (3)). They then define that a subset of predictors $S \subseteq \{1, \dots, d\}$ is *invariant* if $Y - \beta X_S$ is identically distributed across the environments E, where β is the OLS regression coefficient from regressing Y on X_S. The output of ICP is

$$S_{\text{ICP}} \coloneqq \bigcap_{S:\, X_S \text{ is invariant}} S. \tag{4}$$

They show that $S_{\text{ICP}} \subseteq \text{PA}_Y$ (though the inclusion is in some cases strict), meaning that S_{ICP} can be used to learn causal parents. If $\hat{S}_{\text{ICP}}$ is the empirical version of (4) estimated from data, where we use an invariance test with level α instead of ground truth invariance, then $\mathbb{P}(\hat{S}_{\text{ICP}} \subseteq \text{PA}_Y) \geq 1 - \alpha$.

We can understand how ICP exploits exogeneity in the context of Example 3. In **[AncSearch]**, we show that $S_{\text{ICP}} = (\text{PA}_Y \cap \text{CH}_E) \cup (\text{PA}_Y \cap \text{PA}(\text{AN}_Y \cap \text{CH}_E))$. If $X_j \in \text{PA}_Y \cap \text{CH}_E$ (in which case the parenthood to Y can be identified by ICP), then the skeleton of the graph contains $E - X_j - Y$, and using exogeneity of E, one can also use the reasoning in Example 3 to conclude that X_j is a parent of Y. Similarly, for every $X_j \in \text{PA}_Y \cap \text{PA}(\text{AN}_Y \cap \text{CH}_E)$, we could use the skeleton-and-v-structures characterization similar to that of Example 3 to show that the assumption that E is exogenous implies that $X_j \to Y$.

1.4. Causal Models in Distribution Shift

When dealing with a distribution shift, where data is generated from several different distributions, we can use causal models to model those differences. In particular, we may describe these distributions as originating from the same SCM, but subject to different interventions. Though the intervened variable(s) can then change arbitrarily, this encodes the assumption that all remaining parts of the SCM remain unchanged [Haavelmo, 1944, Aldrich, 1989]. In particular, (3) states that the density factorizes into a product of conditionals, where only the intervened conditional(s) change.

This invariance of non-intervened conditionals can be used in various ways. For example, it can be used to draw inference about an intervened distribution using knowledge of the original SCM. This is the case in **[ShiftEval]** and **[ShiftTest]**, where we assume that we know which conditional in an observed distribution Q is changing, and use this to describe properties of the resulting intervened distribution P. Or, reversely, it can be used to infer knowledge of the SCM, when we observe data from various interventions. This is done in **[AncSearch]** and **[InvPolicy]**, where we search for conditionals that do not change across several observed distributions, and use this to infer parts of the underlying causal structure.

Importance Sampling

One way of estimating quantities under distribution shift is to use importance sampling [Horvitz and Thompson, 1952, Shimodaira, 2000]. Importance sampling works by first computing weights $w(x) := p(x)/q(x)$, where p and q are the densities (with respect to the same background measure) of a target distribution P and an observed distribution Q respectively.

If X is a random variable with distribution Q and $T(X)$ is any (reasonably well-behaved) function, then

$$\mathbb{E}_Q[w(X) \cdot T(X)] = \int (p(x)/q(x) \cdot T(x))\, q(x) \mathrm{d}x = \int T(x) p(x) \mathrm{d}x = \mathbb{E}_P[T(X)],$$

where $\mathbb{E}_Q$ and $\mathbb{E}_P$ are expectations in Q and P respectively. This means that if we observe n i.i.d. samples $X_1, \ldots, X_n$ from Q, we can use the law of large numbers to estimate expectations in P by

$$\mathbb{E}_P[T(X)] \approx \frac{1}{n} \sum_{i=1}^{n} w(X_i) T(X_i). \tag{5}$$

When we consider an observed distribution Q and a target distribution $P = Q^{\mathrm{do}(X_j = \tilde{f}(\tilde{\mathrm{PA}}_j, \tilde{\varepsilon}_j))}$ with densities q and p respectively, the factorizations in (2) and (3) imply that q and p factorize in a way where only the conditional densities of X_j given its causal parents differ,

$$w(x) = \frac{p(x)}{q(x)} = \frac{p(x_j | x_{\widetilde{\mathrm{PA}}_j})}{q(x_j | x_{\mathrm{PA}_j})}.$$

This means that in order to compute the importance weights w, we do not need to know the entire distributions P and Q, but only the conditional of the variable that is intervened upon.

In order for the importance sampling weights $w(x) = p(x)/q(x)$ to be well-defined, we require that the support of P is a subset of the support of Q, so the denominator $q(x)$ is non-zero. Even when this is in theory satisfied, importance sampling may face problems in practice, if there are parts of the support of P where $q(x)$ is close to zero, since this can lead to the weights being enormous, and will make the variance of the estimate in (5) be enormous, too.

Since importance sampling enables estimation of means in interventional distributions, we can use importance sampling as an alternative to the covariate adjustment, discussed in Section 1.2. When estimating the effect of a hard intervention $\mathrm{do}(X := x_0)$ on Y in Fig. 2, covariate adjustment requires estimation of the conditional $\mathbb{P}_P(Y = y | C = c, X = x_0)$; an importance sampling approach for the same intervention requires estimation of the conditional $\mathbb{P}_P(X = x_0 | C = c)$, in order to compute weights $w(x, y, c) = \frac{\mathbb{1}_{\{X = x_0\}}}{\mathbb{P}_P(X = x_0 | C = c)}$. Thus, the mean in the intervention distribution can be computed in two different ways; Robins et al. [1994], Robins and Rotnitzky [1995], Cher-

nozhukov et al. [2018] discuss ways of combining these two approaches to create an estimator which is statistically efficient and doubly robust. Here double robustness means that if at least one of the two conditionals is estimated consistently, then the overall procedure is consistent.

2. Shifts in Distributions: Testing

This chapter contains the following paper:

[ShiftTest] [Thams et al., 2021]. N. Thams, S. Saengkyongam, N. Pfister, and J. Peters. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

The paper assumes that we have data from one distribution Q and that we want to test a hypothesis in an unobserved target distribution P. The target distribution P for example can be thought of as the result of 'the world changing' or the data being sampled in environments different from the observed distribution Q; this is typically the motivation in the distribution shift literature [e.g. Quionero-Candela et al., 2009] and in all the papers in Chapter 3. But, as we discuss in this paper, P can also be thought of as the result of some fictional change introduced by the modellers, such as the act of removing confounding between two variables that are confounded in Q.

We formalize the problem of testing under shifts and provide a number of applications of this framework. These include not only cases where the hypothesis of interest is in an unobserved distribution P, but also cases where the hypothesis of interest is in the observed distribution Q, but the hypothesis can more easily be tested by introducing an auxiliary shifted distribution P, and testing the hypothesis there.

Unlike the importance sampling introduced in Section 1.4 in Chapter 1 and the papers in Chapter 3, we do not estimate a mean in the shifted distribution, but instead conduct tests. This implies that weighted-average approaches may not work. To to overcome this, we propose a resampling methodology and provide a number of theoretical results on resampling, which enable guarantees on finite and asymptotic level of tests applied to resampled data.

Statistical Testing under Distributional Shifts

Nikolaj Thams, Sorawit Saengkyongam, Niklas Pfister and Jonas Peters

Abstract

We introduce statistical testing under distributional shifts. We are interested in the hypothesis $P^* \in H_0$ for a target distribution P^*, but observe data from a different distribution Q^*. We assume that P^* is related to Q^* through a known shift τ and formally introduce hypothesis testing in this setting. We propose a general testing procedure that first resamples from the observed data to construct an auxiliary data set and then applies an existing test in the target domain. We prove that if the size of the resample is of order $o(\sqrt{n})$ and the resampling weights are well-behaved, this procedure inherits the pointwise asymptotic level and power from the target test. If the map τ is estimated from data, we maintain the above guarantees under mild conditions on the estimation. Our results extend to finite sample level, uniform asymptotic level and a different resampling scheme, as well as statistical inference different from testing. Testing under distributional shifts allows us to tackle a diverse set of problems. We argue that it may prove useful in contextual bandit problems and covariate shift, we show how it reduces conditional to unconditional independence testing and we provide example applications in causal inference.

1. Introduction

Testing scientific hypotheses about an observed data generating mechanism is an important part of many areas of empirical research and is relevant for almost all types of data. In statistics, the data generating mechanism is described by a distribution P^* and the process of testing a hypothesis corresponds to testing whether P^* belongs to a subclass of distributions H_0. In practice, observations from P^*, for which we want to test the hypothesis $P^* \in H_0$, may not always be available. For instance, sampling from P^* may be unethical if it corresponds to assigning patients to a certain treatment. P^* may also represent the response to a policy that a government is considering to introduce. Yet, in many cases, one may still have data from a different, but related, distribution Q^*. In

the examples above, this could be data from an observational study or under the policies currently deployed by the government.

Although specialized solutions exist for many such problems, there is no general method for tackling them. In this paper, we aim to analyze the above testing task from a general point of view. We assume that a distributional shift $\tau : \mathcal{Q} \to \mathcal{P}$ between domains $\mathcal{Q}$ and $\mathcal{P}$ is known and, using data from Q^*, aim to test the hypothesis $P^* = \tau(Q^*) \in H_0$. We propose the following general framework. We resample from Q^* to construct an auxiliary data set mimicking a sample from P^* and then exploit the existence of a test in the target domain. We assume that Q^* and P^* are absolutely continuous with respect to the same background product measure[1] . Our method does not assume full knowledge of Q^* or P^*, but only knowledge of the (potentially unnormalized) ratio p^*/q^*, where q^* and p^* are densities of Q^* and P^* with respect to the background measure, respectively. If, for example, the shift corresponds to a change in the conditional distribution of a few of the observed variables, one only needs to know these changing conditionals.

Our framework assumes the existence of a test φ in the target domain, i.e., a test that could be applied if data from P^* were available. This test φ is then applied to a resampled version of the observed data set. Here, we propose a sampling scheme that is similar to sampling importance resampling (SIR), proposed by Rubin [1987] and Smith and Gelfand [1992] but generates a distinct sample of size m, using weights $r(X) \propto q^*(X)/p^*(X)$ where X is a random vector with distribution Q^*. We prove that this procedure inherits the pointwise asymptotic properties of the test φ if the weights r have finite second moment in Q^*, and $m = o(\sqrt{n})$. In particular, the procedure holds pointwise asymptotic level if the test φ does. We show that the same can be obtained if r is not known, but can be estimated from data sufficiently well. The proposed method is easy-to-use and can be applied to any hypothesis test, even if the test is based on a nonlinear test statistic.

Several problems can be cast as hypothesis tests under distributional shifts. This includes hypothesis tests in off-policy evaluation, tests of conditional independence, testing the absence of causal edges through dormant independences [Verma and Pearl, 1991, Shpitser and Pearl, 2008], that is, testing certain equality constraints in an observed distribution, and problems of covariate shift. Our proposed method can be applied to all of these problems. For some of them, we are not aware of other methods with theoretical guarantees – this includes dormant independence testing with continuous variables, off-policy testing with complex hypotheses, conditional independence testing with censored data and model selection under covariate shift with complex scoring functions. The framework also inspired a novel method for causal discovery that exploits knowledge of a single causal conditional. For some of the above problems, however, more specialized solutions exist, and as such, the proposed testing procedure relates to a line of related work.

Ratios of densities have been applied in the reinforcement learning literature [e.g.

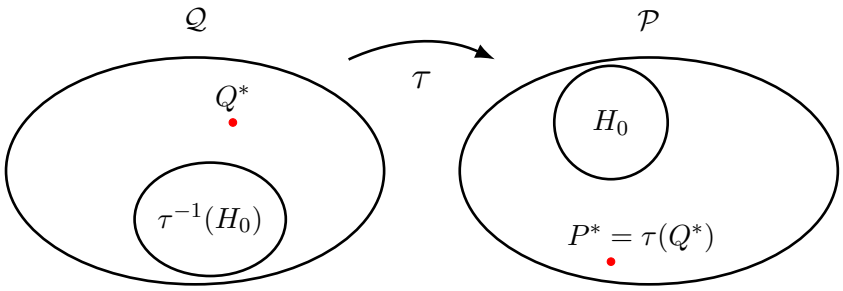
[1] We believe that most of the statements still hold if Q^* and P^* are absolutely continuous with respect to the same non-product measure but make the assumption of product measures for simplicity.

Sutton and Barto, 1998], where inference in P^* using data from Q^* is known as off-policy prediction [Precup et al., 2001]. One can estimate the expectation of X under P^* using importance sampling, (IS), that is, as weighted averages using weights $r(X)$, possibly truncated to decrease the variance of the estimation [Precup et al., 2001, Mahmood et al., 2014]. An approach to estimate means in P^* based on resampling was proposed by Schlegel et al. [2019]. However, they consider the size of the resample fixed and do not consider statistical testing. Thomas et al. [2015] propose bootstrap confidence intervals for off-policy prediction based on important weighted returns. Hao et al. [2021] present a bootstrapping approach with Fitted Q-Evaluation (FQE) for off-policy statistical inference and demonstrate its distributional consistency guarantee.

In the causal inference literature, inverse probability weighting (IPW) To estimate the effect of a treatment X on a response Y, one can weight each observed response Y_i with $1/q^*(X_i|Z_i)$, where Z_i is an observed confounder. For continuous treatments, it has been proposed to change the numerator to a marginal distribution $p^*(x)$ to stabilize the weights [Hernán and Robins, 2006, Naimi et al., 2014]. Both choices of weights appear in our framework, too (e.g., the first one corresponds to a target distribution with $p^*(x|z) \propto 1$). In general, IS and IPW can only be applied if the population version of the test statistic can be written as a mean of a function of a single observation, such as $\mathbb{E}[Y_i]$ or $\mathbb{E}[f(X_i, Z_i)]$, whereas our approach also applies to test statistics that are functions of the entire sample, which is the case for many tests that go beyond testing moments, such as several independence tests, for example.

SIR sampling schemes were first studied by Rubin [1987] and are often used in the context of Bayesian inference [Smith and Gelfand, 1992]. Skare et al. [2003] show that when using weighted resampling with or without replacement, for $n \to \infty$ and fixed m, the sample converges towards m i.i.d. draws from the target distribution, and provide rates for the convergence. Our work is inspired by these types of results, even though our proofs require different techniques.

Our paper adds to the literature on distributional shifts by considering hypothesis tests in shifted distributions. In the context of prediction, distributional shifts, or dataset shifts, have been studied in the machine learning literature both to handle the situations where a marginal covariate distribution changes and when the conditional distribution of label given covariate changes [Quionero-Candela et al., 2009]. If the shift represents a changing marginal distribution and unlabelled samples are available from both training and test environments, Huang et al. [2006] propose kernel mean matching, which non-parameterically reweights the training loss to resemble the loss on a target sample. In settings where a generative model and causal graph is known, Pearl and Bareinboim [2011], Subbaswamy et al. [2019] provide graphical criteria under which causal estimates can be 'transported' from one distribution Q to a shifted distribution P, assuming knowledge of both joint distributions Q and P. In contrast, we consider statistical testing, and neither assume knowledge of the full causal graph nor availability of samples from the target distribution, but instead knowledge of how the target data differ from the observed data. Conformal prediction [Vovk et al., 2005] has been applied to covariate shift [Tibshirani et al., 2019, Park et al., 2021] too, but its goal of constructing a prediction interval for a new random observation is generally different than the one of



Distributions on observed domain $\mathcal{X}$; $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q^*$.

Distributions on target domain $\mathcal{Z}$; unobserved.

Figure 1.: Illustration of observed and target domains, $\mathcal{Q}$ and $\mathcal{P}$, target hypothesis $H_0 \subseteq \mathcal{P}$ and pullback hypothesis $\tau^{-1}(H_0)$.

the proposed testing framework (see, however, the extension of our theory to general inference problems, Section 4.6).

This paper contains four main contributions: First, we formally define testing under distributional shifts, and define notions such as pointwise and asymptotic level when using observed data to test the hypothesis in the target domain. Second, we outline a number of statistical problems, that can be solved by testing under a shift, including conditional independence testing and testing dormant independences. Third, we propose methods that enable testing under distributional shifts: both a simple method based on rejection sampling and a resampling scheme that requires fewer assumptions than the rejection sampler. Fourth, we provide finite sample and asymptotic guarantees for our proposed resampling scheme; contrary to the existing literature, where typically m fixed and $n \to \infty$ has been studied [e.g., Skare et al., 2003], we study the asymptotic behaviour of our resampling test when both m and n approach infinity, and show that under any resampling scheme, the requirement $m = o(\sqrt{n})$ is necessary.

2. Statistical Testing under Distributional Shifts

2.1. Testing hypotheses in a target distribution

Consider a set of distributions $\mathcal{P}$ on a target domain $\mathcal{Z} \subseteq \mathbb{R}^d$ and a null hypothesis $H_0 \subseteq \mathcal{P}$. In hypothesis testing, we are usually given data from a distribution $P^* \in \mathcal{P}$ and want to test whether $P^* \in H_0$. In this paper, we consider the problem of testing the same hypothesis but instead of observing data from P^* directly, we assume the data are generated by a different, but related, distribution Q^* from a set of distributions $\mathcal{Q}$ over a (potentially) different observational domain $\mathcal{X} \subseteq \mathbb{R}^e$.

More formally, we assume that we have observed data $\mathbf{X}_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$ consisting of n i.i.d. random variables X_i with distribution $Q^* \in \mathcal{Q}$. We use superscripts to denote the individual coordinates of $X = (X^1, \ldots, X^d) \in \mathcal{X}$. We assume that Q^* and

P^* are related through a map $\tau : \mathcal{Q} \to \mathcal{P}$, called a (distributional) shift, which satisfies $P^* = \tau(Q^*)$. We aim to construct a randomized hypothesis test $\psi_n : \mathcal{X}^n \times \mathbb{R} \to \{0, 1\}$ that we can apply to the observed data $\mathbf{X}_n$ to test the null hypothesis

$$\tau(Q^*) \in H_0. \tag{1}$$

We reject this null hypothesis if $\psi_n = 1$ and do not reject the null if $\psi_n = 0$. To allow for random components, we let ψ_n take as input a uniformly distributed random variable U (assumed to be independent of the other variables) that generates the randomness of ψ_n. Whenever there is no ambiguity about the randomization, we omit U and write $\psi_n(\mathbf{X}_n)$; unless stated otherwise, any expectation or probability includes the randomness of U. For $\alpha \in (0, 1)$, we say that ψ_n holds level α at sample size n if it holds that

$$\sup_{Q \in \tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{2}$$

In practice, requiring level at sample size n is often too restrictive. We say that the test has pointwise asymptotic level α if

$$\sup_{Q \in \tau^{-1}(H_0)} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{3}$$

We illustrate the setup in Fig. 1.

Remark 1. The map $\tau : \mathcal{Q} \to \mathcal{P}$ above represents a view that starts with the distribution Q^* of the observed data and considers the distribution P^* of interest as the image under τ. Alternatively, one may also start with a map $\eta : \mathcal{P} \to \mathcal{Q}$ and say that the test holds level α at sample size n if

$$\sup_{P \in H_0} \mathbb{P}_{\eta(P)}(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{4}$$

This corresponds to a level guarantee for a test of the hypothesis $\eta^{-1}(Q^*) \cap H_0 \neq \emptyset$. If τ is invertible, the two views trivially coincide with $\eta := \tau^{-1}$, but in general there are subtle differences, see Appendix A.1.1 for details.

2.2. Distributional shifts

We consider two types of maps $\tau : \mathcal{Q} \to \mathcal{P}$, both of which can be written in product form. First, assume that there is a subset $A \subseteq \{1, \dots, d\}$ together with a known map $r : x^A \mapsto r(x^A) \in [0, \infty)$ such that for all $q \in \mathcal{Q}$ the target density[2] $\tau(q)$ satisfies that

$$\tau(q)(x^1, \dots, x^d) \propto r(x^A) \cdot q(x^1, \dots, x^d) \qquad \text{for all } (x^1, \dots, x^d) \in \mathcal{Z}. \tag{5}$$

[2] In the remainder of this work, we assume that $\mathcal{X}$ and $\mathcal{Z}$ are both subsets of $\mathbb{R}^d$, that is $e = d$, and that all distributions in $\mathcal{P}$ and $\mathcal{Q}$ have densities with respect to the same dominating product measure μ. We refer to a distribution Q and its density q interchangeably.

Here, we assume that the factor r is known in the sense that it can be evaluated for any given x^A (or at least on all points in the observed sample $\mathbf{X}_n^A$). This type of map naturally arises in many examples, such as in off-policy evaluations with a known training policy or when performing a conditional independence test with known conditional, see Section 3.2.

Second, assume that there is a subset $A \subseteq \{1, \dots, d\}$ together with a known map $r_{(\cdot)} : (q, x^A) \mapsto r_q(x^A) \in [0, \infty)$ such that for all $q \in \mathcal{Q}$, the target density $\tau(q)$ satisfies that

$$\tau(q)(x^1, \dots, x^d) \propto r_q(x^A) \cdot q(x^1, \dots, x^d) \qquad \text{for all } (x^1, \dots, x^d) \in \mathcal{Z}. \tag{6}$$

Here, we assume that the factor $r_{(\cdot)}$ can be evaluated for any given (q, x^A). This case arises, for example, when the training policy or the conditional is unknown and needs to be estimated from data. If, in any of the above two cases, the set A is not mentioned explicitly, we implicitly assume $A = \{1, \dots, d\}$. In many applications τ represents a local change in the system, so even though d may be large, $|A|$ will be much smaller than d. In particular we do not need to know the entire distribution to evaluate $r(x^A)$.

In principle, this approach applies to any full-support distribution Q, since for a given target distribution $P \in \mathcal{P}$, $P = \tau(Q)$ is satisfied as long as we define $r(x) = p(x)/q(x)$, and in the case that we consider a change in a single conditional, this simplifies to $r(x^{\{a_1,a_2\}}) = p(x^{a_1}|x^{a_2})/q(x^{a_1}|x^{a_2})$. In practice, there may be regions of the support where $q(x)$ is much smaller than $p(x)$, in which case the weights will be ill-behaved. We address this issue in Assumption (A2) and analyze its impact in Theorem 4. For some shifts and hypotheses, direct solutions are available that do not use the importance weights (5): For example, when testing a hypothesis about X^1 under a mean shift in the marginal distribution of X^1, one could directly add the anticipated shift in mean to every observation before testing. However, in most cases involving shifts in conditional distributions or in variables different from those entering the test, such approaches fail. If τ is misspecified, in the sense that $\tau(Q^*) \neq P^*$, then the guarantees for the methodology below still hold, but for testing the distribution $\tau(Q^*) \in H_0$ instead of $P^* \in H_0$.

2.3. Exploiting a test in the target domain

In this work, we assume that there is a test φ for the hypothesis H_0 that can be applied to data from the target domain $\mathcal{Z}$. Formally, we consider a sequence $\varphi_k : \mathcal{Z}^k \times \mathbb{R} \to \mathbb{R}$ of (potentially randomized) hypothesis tests for H_0 that can be applied to k observations $\mathbf{Z}_k$ from the target domain $\mathcal{Z}$ and a uniformly distributed random variable V, generating the randomness of φ_k. For simplicity, we omit V from the notation and write $\varphi_k(\mathbf{Z}_k)$. We say that $\varphi := (\varphi_k)_k$ has pointwise asymptotic level α for H_0 in the target domain if

$$\sup_{P \in H_0} \limsup_{k \to \infty} \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1) \leq \alpha. \tag{7}$$

To address the problem of testing under distributional shifts, we propose in Section 4 to resample a data set of size m from the observed data $\mathbf{X}_n$ (using resampling weights that

depend on the shift) and apply the test φ_m to the resampled data. This procedure is easy-to-use and can be combined with any testing procedure φ from the target domain.

2.4. Testing hypotheses in the observed domain

The framework of testing hypotheses in the target distribution can be helpful even if we are interested in testing a hypothesis about the observed distribution Q^*, that is, testing $Q^* \in H_0^{\mathcal{Q}}$ for some $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$. If $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}} := H_0$, any test ψ_n satisfying pointwise asymptotic level (3) for $H_0^{\mathcal{P}} \subseteq \mathcal{P}$ can be used as a test for $Q^* \in H_0^{\mathcal{Q}}$, and will still satisfy asymptotic level, see Section 4.4.5.

Such an approach can be particularly interesting when it is more difficult to test $Q^* \in H_0^{\mathcal{Q}}$ in the observed domain than it is to test $\tau(Q^*) \in H_0^{\mathcal{P}}$ in the target domain. For example, testing conditional independence in the observed domain can be reduced to (unconditional) independence testing in the target domain. Here, we may benefit from transferring the test into the target domain if one of the conditionals is known or can be estimated from data. Also testing a Verma equality [Verma and Pearl, 1991] in the observed distribution can be turned into an independence test in the target distribution, too; but here, testing directly in the observed domain may not even be possible. Often there is a computational advantage of our approach: In many situations, the resampled data set, where the hypothesis is easier to test, is much smaller than the original data set, see for instance the experiment in Section 5.4. When the hypothesis of interest is in the observed domain, usually different choices for the target distribution are possible. In practice, it is helpful to choose a target distribution that yields well-behaved resampling weights (see (10)), which can often be achieved by matching certain marginals, see, e.g., Section 5.6 [see also Robins et al., 2000, Hernán and Robins, 2006].

The following Section 3 discusses the above and other applications of testing under distributional shifts in more detail. Corresponding simulation experiments are presented in Section 5. Section 4 provides details of our method and its theoretical guarantees.

3. Example Applications of Testing under Distributional Dhifts

3.1. Conditional independence testing

Let us first consider a random vector (X, Y, Z) with joint probability density function q^* and assume that the conditional $q^*(z|x)$ is known. We can then apply our framework to test

$$H_0^{\mathcal{Q}} = \{Q : X \perp\!\!\!\perp Y \mid Z \text{ and } q(z|x) = q^*(z|x)\} \quad \text{(cond. ind. in observed domain)}$$

by reducing the problem to an unconditional independence test. The key idea is to factor a density $q \in H_0^{\mathcal{Q}}$ as $q(x, y, z) = q(y|x, z)q^*(z|x)q(x)$, replace[3] the conditional $q^*(z|x)$

[3] If the factorization happens to correspond to the factorization using a causal graph, this is similar to performing an intervention on Z, see Appendix A.1.2. However, the proposed factorization is always valid, so this procedure does not make any assumptions about causal structures.

by, e.g., a standard normal density $\phi(z)$ to obtain the target density p, and then test for unconditional independence of X and Y. When X is a randomized treatment, Y the outcome, and Z is a mediator, this corresponds to testing (non-parametrically) the existence of a direct causal effect [e.g., Pearl, 2009, Imbens and Rubin, 2015, Hernán and Robins, 2020].

Formally, we define a corresponding hypothesis in the target domain:

$$H_0^{\mathcal{P}} := \{P : X \perp\!\!\!\perp Y \text{ and } p(z|x) = \phi(z)\} \quad \text{(ind. in target domain)}$$

with ϕ being the standard normal density. We can then define a map τ by

$$\tau(q)(x, y, z) := \frac{\phi(z)}{q^*(z|x)} \cdot q(x, y, z) \qquad \text{for all } (x, y, z) \in \mathcal{Z}.$$

Considering any $q \in H_0^{\mathcal{Q}}$ and writing $p := \tau(q)$, we have

$$p(x, y, z) = \frac{\phi(z)}{q^*(z|x)} q(y|x, z) q^*(z|x) q(x) = q(y|x, z) q(x) \phi(z).$$

This shows[4] that conditional independence $X \perp\!\!\!\perp Y \mid Z$ in Q implies independence $X \perp\!\!\!\perp Y$ in P and therefore $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$. Starting with an independence test φ_m for $H_0^{\mathcal{P}}$, we can thus test $\tau(Q^*) \in H_0^{\mathcal{P}}$, with level guarantee in (3). As we have argued in Section 2.4, this corresponds to testing $Q^* \in H_0^{\mathcal{Q}}$, and thereby reduces the question of conditional independence to independence.

If, instead of $q^*(z|x)$, we know the reverse conditional $q^*(x|z)$, we can use the same reasoning as above using the factorization $q(x, y, z) = q(z) q^*(x|z) q(y|x, z)$ and a marginal target density $\phi(x)$ to again test $X \perp\!\!\!\perp Y \mid Z$. When X is a treatment, Y the outcome, Z is the full set of covariates, and $q^*(x|z)$ represents the randomization scheme, this corresponds to testing (non-parametrically) the existence of a total causal effect [e.g., Peters et al., 2017] between X and Y.

If neither of the conditionals is known, we can still fit the test into our framework. To do so, define the hypotheses $H_0^{\mathcal{Q}} := \{Q : X \perp\!\!\!\perp Y | Z\}$, $H_0^{\mathcal{P}} := \{P : X \perp\!\!\!\perp Y \text{ and } p(z|x) = \phi(z)\}$, and the map τ via $\tau(q)(x, y, z) := \frac{\phi(z)}{q(z|x)} \cdot q(x, y, z)$, for all $(x^1, \ldots, x^d) \in \mathcal{Z}$; estimate the conditional $q(z|x)$ from data and may still maintain the level guarantee of the overall procedure. There are other, more specialized conditional independence tests but this viewpoint may be an interesting alternative if we can estimate one of the conditionals well, e.g., because there are many more observations of (X, Z) than there are of (X, Z, Y). Furthermore, conditional independence tests may not even be available in some complex settings, while marginal independence tests may exist. For example, we illustrate in Section 5.5 that our method can be applied to conduct a non-parametric test for conditional independence with right-censored data. To the best of our knowledge, there are no other non-parametric conditional independence tests available

[4] The following statement holds because, clearly, $p(z|x) = \phi(z)$ and if $X \perp\!\!\!\perp Y \mid Z$ in q, that is, $q(y|x, z) = q(y|z)$ for all x, y, z yielding this expression well-defined, it follows $p(x, y) = p(x)p(y)$.

in this setting.

The assumption of knowing one conditional is also exploited by the conditional randomization (CRT) and the conditional permutation test (CPT) by Candès et al. [2018] and Berrett et al. [2020], respectively. These methods, however, require knowledge of $q^*(x|z)$ and cannot exploit knowledge of $q^*(z|x)$. They simulate (in case of CRT) or permute (in case of CPT) X while keeping Z and Y fixed and construct p-values for the hypothesis of conditional independence. The approaches are similar in that they use the known conditional to create weights. Our method, however, explicitly constructs a target distribution and, as argued above, cannot only exploit knowledge of $q^*(x|z)$ but also knowledge of $q^*(z|x)$.

Our approach can be modified to obtain a double robustness property. Suppose we consider a map $\tau(q)(x,y,z) = q(x,y,z) \cdot \frac{\phi_X(x)}{q(x|z)} \frac{\phi_Y(y)}{q(y|z)}$, for some density functions $\phi_X(x)$ and $\phi_Y(y)$ and model conditionals $q(x|z)$ and $q(y|z)$. Then

$$\tau(q^*)(x,y,z) = q^*(z)q^*(x,y|z)\frac{\phi_X(x)\phi_Y(y)}{q(x|z)q(y|z)} = q^*(z)\phi_X(x)\phi_Y(y)\frac{q^*(x,y|z)}{q(x|z)q(y|z)}.$$

Suppose, for example, that $q(x|z) = q^*(x|z)$, but $q(y|z) \neq q^*(y|z)$. Then, it still holds under the null hypothesis $X \perp\!\!\!\perp Y|Z$ in Q^* that $\tau(q^*)(x,y,z) = \phi_X(x)q^*(z)\phi_Y(y)\frac{q^*(y|z)}{q(y|z)}$, meaning that $X \perp\!\!\!\perp Y$ in $\tau(Q^*)$. Similarly, if $q(y|z) = q^*(y|z)$, but $q(x|z) \neq q^*(x|z)$, conditional independence in Q^* also implies independence in $\tau(Q^*)$. That is, as long as one of the modelled conditionals $q(x|z)$ and $q(y|z)$ equals the corresponding one of q^* (we do not need to know which one), the hypothesis of conditional independence in Q^* can be tested as a hypothesis of marginal independence in P^*. This is similar to the doubly robustness guarantee in [Shi et al., 2021], where as long as one estimates at least one conditional consistently, the overall test is consistent.

3.2. Off-policy testing

Consider a contextual bandit setup [e.g. Langford and Zhang, 2008, Agarwal et al., 2014]. In each round, an agent observes a context $Z := (Z^1, \dots, Z^d)$ and selects an action $A \in \{a_1, \dots, a_L\}$, based on a known policy $q^*(a|z)$. The agent then receives a reward R depending on the chosen action A and the observed context Z. Suppose we have access to a data set $\mathbf{X}_n$ of n rounds containing observations $X_i := (Z_i, A_i, R_i)$, $i = 1, \dots, n$. We can then test statements about the distribution under another policy $p^*(a|z)$. For example, we can test whether the expected reward is smaller than zero. To do so, we define

$$H_0 := \{P : \mathbb{E}_P[R] \leq 0 \quad \text{and} \quad p(a|z) = p^*(a|z)\}$$

and $\tau(q)(x) := r(x)q(x)$ with the shift factor $r(z,a) := p^*(a|z)/q^*(a|z)$. Here, the function of interest can be written as an expectation of a single observation, so other, simpler approaches such as IS or IPW can be used, too (see Section 1).

But it is also possible to test more involved hypotheses. This includes testing (condi-

tional) independence under a new policy, for example. Suppose that one of the covariates Z^j is used for selecting actions by an observed policy $q^*(a|z)$. This creates a dependence between Z^j and R, but it is unclear whether this dependence is only due to the action A being based on Z^j, or whether Z^j also depends on R in other ways, for instance in that Z^j has a direct effect on R. To test the latter statement, we can create a new policy $p^*(a|z)$ that does not use Z^j for selecting actions. Then, we can test whether, under $p^*(a|z)$, R is independent of Z^j, given the other variables that the action is based on. If not, we know that there must be a dependence between R and Z^j under $q^*(a|z)$ beyond the action A being based on Z^j. This may be relevant for learning sets of features that are invariant across different environments, that is, features Z^J such that $R \mid Z^J$ is stable across environments. A policy that depends on such invariant features is guaranteed to generalize to unseen environments [Saengkyongam et al., 2021]. Another, more involved hypothesis for off-policy evaluation compares the reward distributions under two different policies. This can be written as a two-sample test, which we discuss in Section 3.3.

This procedure extends to more general reinforcement learning settings, where for example one repeatedly observes a Markov decision process [Sutton and Barto, 1998]. The weights then correspond to a product containing one factor for each decision. If the decision process contains many decisions or the data generating policy is not sufficiently close to the policy to be evaluated, off-policy evaluation becomes a difficult problem, with weights not being well-behaved. This problem is well-known in the reinforcement literature [Mahmood et al., 2014, Levine et al., 2020], where the ill-behaved weights result in large variance of estimators; in the methodology we propose in Section 4, it generally results in a loss of power.

3.3. Two-sample testing with one transformed sample

We can use the framework to perform a two-sample test, after transforming one of the two samples. Consider the observed distribution q^* over $X = (X^1, \dots, X^d) \in \mathbb{R}^d$ and $K \in \{1, 2\}$, where the latter indicates which of the two samples a data point belongs to. We now keep the first sample as it is and change the second sample with a transformation τ, $q^* \mapsto \tau(q^*)$. We can then test whether, after the transformation, the two samples come from the same distribution, i.e., whether

$$q^*(x|k = 1) = \tau(q^*)(x|k = 2)$$

for all x. For example, assume that we know the conditional $q^*(x^1|x^2, k = 2)$ and consider transforming this to $p^*(x^1|x^2, k = 2)$. To formally apply our framework, we then define

$$H_0 := \{P : (X^1, \dots, X^d)_{|K=1} \overset{\mathcal{L}}{=} (X^1, \dots, X^d)_{|K=2}, \quad p(x^1|x^2, k = 2) = p^*(x^1|x^2, k = 2)\}$$

and the shift $\tau(q)(x^1, \dots, x^d, k) := r(x^1, x^2, k) \cdot q(x^1, \dots, x^d, k)$, where

$$r(x^1, x^2, k) := \begin{cases} 1 & \text{if } k = 1 \\ \frac{p^*(x^1|x^2,k=2)}{q^*(x^1|x^2,k=2)} & \text{if } k = 2. \end{cases}$$

In particular this approach can be used for off-policy evaluation (the setting described in the previous section) to test whether the reward under the training policy $q^*(a|z)$ has the same distribution as under a target policy $p^*(a|z)$. We first randomly split the training sample into two subsamples ($K = 1$ and $K = 2$) and then test whether the distribution of the reward is different under the two policies,

$$H_0 := \{P : R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2},\ p(a|z, k = 1) = q^*(a|z) \text{ and } p(a|z, k = 2) = p^*(a|z)\},$$

by using weights $r(a, z, k) = p^*(a|z)/q^*(a|z)$ when $k = 2$ and $r(a, z, k) = 1$ otherwise. This is not confined to testing identical distributions: For example, we can also test, non-parametrically, whether the expected reward under the new policy $p^*(a|z)$ is larger than under the current policy $q^*(a|z)$. To do so, we define

$$H_0 := \{P : \mathbb{E}_P[R_{|K=2}] \leq \mathbb{E}_P[R_{|K=1}],\ p(a|z, k = 1) = q^*(a|z) \text{ and } p(a|z, k = 2) = p^*(a|z)\}.$$

Section 5.3 shows some empirical evaluations of such tests.

3.4. Dormant independences

Let us consider a random vector $(X^1, \dots, X^d)$ with a distribution Q that is Markovian with respect to a directed acyclic graph and that has a density w.r.t. a product measure. By the global Markov condition [e.g. Lauritzen, 1996], we then have for all disjoint subsets $A, B, C \subset \{1, \dots, d\}$ that $X^A \perp\!\!\!\perp X^B \,|\, X^C$ if A d-separates[5] B given C. If some of the components of the random vector are unobserved, the Markov assumption still implies conditional independence statements in the observational distribution. In addition, however, it may impose constraints on the observational distribution that are different from conditional independence constraints. Figure 2 shows a famous example, due to Verma and Pearl [1991], that gives rise to the Verma-constraint: If the random vector (X^1, X^2, X^3, X^4, H) has a distribution Q that is Markovian w.r.t. the graph $\mathcal{G}$ shown in Fig. 2 (left), there exists a function f such that, for all x^1, x^3, x^4,

$$\int_{-\infty}^{\infty} q(x^2|x^1)q(x^4|x^1, x^2, x^3)\, dx^2 = f(x^3, x^4) \tag{8}$$

(in particular, f does not depend on x^1). This constraint cannot be written as a conditional independence constraint in the observational distribution Q. In general, the constraint (8) does not hold if Q is Markovian w.r.t. $\mathcal{H}$ (see Fig. 2, right). Assume now that the conditional $q(x^3|x^2) = q^*(x^3|x^2)$ is known (e.g., through a randomiza-

[5] Whether a d-separation statement holds is entirely determined from the graph; the precise definition of d-separation can be found in [e.g., Spirtes et al., 2000] but is not important here.

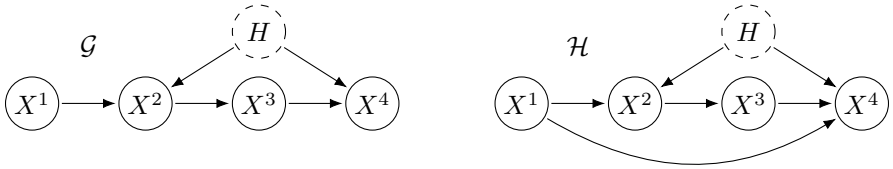


Figure 2.: If Q is Markovian w.r.t. graph $\mathcal{G}$ (left), then Q satisfies the Verma constraint (8). In general, this constraint does not hold if Q is Markovian w.r.t. $\mathcal{H}$ (right). Such constraints can be tested for using the framework of statistical testing under distributional shifts, see Section 3.4.

tion experiment). We can then hope to test for this constraint by considering the null hypothesis

$$H_0^{\mathcal{Q}} := \{Q : Q \text{ satisfies (8) and } q(x^3|x^2) = q^*(x^3|x^2)\}$$

and hence distinguish between $\mathcal{G}$ and $\mathcal{H}$. Constraints of the form (8) have been studied recently, and in a few special cases, such as binary or Gaussian data, the constraints can be exploited to construct score-based structure learning methodology [Shpitser et al., 2012, Nowzohour et al., 2017]. Shpitser and Pearl [2008] show that some of such constraints, called dormant independence constraints, can be written as a conditional independence constraint in an interventional distribution [see also Robins, 1999, Shpitser et al., 2014, Richardson et al., 2017], and Shpitser et al. [2009] propose an algorithm that detects constraints that arise due to dormant independences using oracle knowledge. The Verma constraint (8), too, is a dormant independence, that is, we have

$$X^1 \perp\!\!\!\perp X^4 \qquad \text{in } Q^{do(X^3:=N)}, \tag{9}$$

where $N \sim \mathcal{N}(0,1)$, for example. Here, $Q^{do(X^3:=N)}$, denotes the distribution in which $q^*(x^3|x^2)$ is replaced by $\phi(x^3)$ see Appendix A.1.2 for details. Using the described framework, we can test (9) to distinguish between $\mathcal{G}$ and $\mathcal{H}$.

In practice, we may need to estimate the corresponding conditional, such as $q(x^3|x^2)$ in the example above, from data; as before, this still fits into the framework using (6), see Section 5.6 for a simulation study. In special cases, such as binary, applying resampling methodology to this type of problem has been considered before [Bhattacharya, 2019], but we are not aware of any work proposing a general testing procedure with theoretical guarantees.

The problem of testing (conditional) independences under an interventional distribution has been shown to be relevant in real-world applications. One application is testing direct effects in dynamic treatment regimes [e.g., Robins and Wasserman, 1997]. Consider a two-stage dynamic treatment regime consisting of a sequence of two treatment variables A^0, A^1, an intermediate covariate vector Z^1 (corresponding to X^1, X^3, and X^2 in Fig. 2, respectively), and an outcome variable Y. To test whether there is a direct causal effect of the first treatment A^0 on the outcome Y, one needs to test whether A^0 is independent of Y in the interventional distribution in which the conditional $q^*(z^1|a^1)$ is

replaced by some marginal $\phi(z^1)$. Here, one cannot disentangle the direct causal effect of A^0 on Y and the total causal effect based solely on the (conditional) independence constraints in the observational distribution; the problem is related to the g-null paradox [Robins and Wasserman, 1997, McGrath et al., 2022].

Another application is testing front-door assumptions in causal effect estimation, which has been proposed by Bhattacharya and Nabi [2022]. The front-door adjustment [Pearl, 2009] is an adjustment strategy for identifying the causal effect of a treatment X on an outcome Y in the presence of hidden confounders U. The main assumption is that we observe an intermediate variable M that mediates the effect of X on Y and is not confounded by U. Bhattacharya and Nabi [2022] consider the setup in which we additionally observe an anchor variable Z which causes both the treatment X and the mediator M, but does not directly cause the outcome Y. The anchor variable allows for testing of the front-door assumptions by testing independence constraints under the interventional distribution in which the conditional $q^*(m|x,z)$ is replaced by some marginal $\phi(m)$. We follow Bhattacharya and Nabi [2022] and conduct a numerical experiment on the Framingham heart study dataset [Dawber et al., 1951] in Section 5.7.

3.5. Uncovering heterogeneity for causal discovery

For a response variable Y, consider the problem of finding the causal predictors X^{PA_Y}, with $\mathrm{PA}_Y \subseteq \{1,\dots,d\}$, among a set of potential predictors $X^1,\dots,X^d$. assumes that data are observed in different environments and that the causal mechanism for Y, given its causal predictors PA_Y is invariant over the observed environments [see also Haavelmo, 1944, Aldrich, 1989, Pearl, 2009]. This allows for the following procedure: For all subsets $S \subseteq \{1,\dots,d\}$ one tests whether the conditional $Y|X^S$ is invariant. The hypothesis is true for the set of causal parents, so taking the intersection over all such invariant sets yields, with large probability, a subset of PA_Y Environments can, for example, correspond to different interventions on a node X^j. Using the concept of testing under distributional shifts, we can apply a similar reasoning even if no environments are available and one causal conditional is known instead.

Assume a causal model (e.g., a structural causal model, SCM, see Appendix A.1.2) over the variables $Y, X^1,\dots,X^d$ and denote the causal predictors of X^j by PA_j. Assume further that there is a j for which the conditional $q^*(x^j|x^{\mathrm{PA}_j})$ is known. To infer the causal parents of Y, we now construct a new distribution, in which the conditional $q^*(x^j|x^{\mathrm{PA}_j})$ has been changed to another conditional $p^*(x^j|x^{\mathrm{PA}_j})$ – this corresponds to a distribution generated by an intervention on X^j. We then take the original and the resampled data as two 'environments' and apply the ICP methodology by testing whether the conditional $Y \mid X^S$ is invariant w.r.t. these two environments. That is, in the absence of 'true heterogeneity', we use the known conditional to artificially sample heterogeneity. Formally, for a candidate set $S \subseteq \{1,\dots,d\}$ and an indicator variable K

indexing the two environments, we define the hypothesis

$$H_{0,S} := \{P : Y \mid X^S{}_{|K=1} \overset{\mathcal{L}}{=} Y \mid X^S{}_{|K=2}, \quad p(x^j|x^{\mathrm{PA}_j}, k=1) = q^*(x^{\mathrm{PA}_j}|x^{\mathrm{PA}_j}) \text{ and } p(x^j|x^{\mathrm{PA}_j}, k=2) = p^*(x^{\mathrm{PA}_j}|x^{\mathrm{PA}_j})\}$$

and the shift factor $r(x^j, x^{\mathrm{PA}_j}, k)$ similar to the one in Section 3.3. Naturally, the procedure extends to $K > 2$. The distributional shift corresponds to an intervention on X^j and it follows by modularity[6] that H_{0,PA_Y} is true. Therefore, the intersection over all sets for which $H_{0,S}$ holds trivially satisfies

$$\bigcap_{S:H_{0,S} \text{ holds}} S \subseteq \mathrm{PA}_Y,$$

where we define the intersection over an empty index set as the empty set. Our framework allows for testing such hypotheses from finitely many data (that were generated only using the conditional $q^*(x^j|x^{\mathrm{PA}_j})$) and prove theoretical results that imply level statements for testing $H_{0,S}$. Such guarantees carry over to coverage statements for $\hat{S} := \cap_{S:H_{0,S} \text{ not rej.}} S$, that is, $\hat{S} \subseteq \mathrm{PA}_Y$ with large probability.

3.6. Model selection under covariate shift

Consider the problem of comparing models in a supervised learning task when the covariate distribution changes compared to the distribution that generated the training data. Formally, let us consider an i.i.d. sample $D := \{(X_i, Y_i)\}_{i=1}^n$ from a distribution q^*, where $X_i \in \mathcal{X}$ are covariates with density $q^*(x)$ and $Y_i \in \mathcal{Y}$ is a label with conditional density $q^*(y|x)$. First, we randomly split the sample into two distinct sets, which we call training set D_{train} and test set D_{test}. Let $\hat{f}_1 : \mathcal{X} \to \mathcal{Y}$ and $\hat{f}_2 : \mathcal{X} \to \mathcal{Y}$ be outputs of two supervised learning algorithms trained on D_{train}. In model selection under covariate shift [e.g. Quionero-Candela et al., 2009], we are interested in comparing the performance of the predictors $\hat{f}_1$ and $\hat{f}_2$ on a distribution p^*, where the covariate distribution is changed from $q^*(x)$ to $p^*(x)$, but the conditional $p^*(y|x) = q^*(y|x)$ remains the same. If we had an i.i.d. data set D_{test}^{sh} from the shifted distribution p^*, we could compare the performances using a scoring function $\mathcal{S}(D_{test}^{sh}, \hat{f})$ that for each of the predictors outputs a real-valued evaluation score. However, we only have access to D_{test}, which comes from q^*. Let us for now assume that the shift from $q^*(x)$ to $p^*(x)$ is known. Existing methods use IPW to correct for the distributional shift [Sugiyama et al., 2007], which requires that the scoring function can be expressed in terms of an expectation of a single observation, such as the mean squared error. However, such a decomposition is not immediate for many scoring functions as for example the area under the curve (AUC). The framework of testing under distributional shifts allows for an arbitrary scoring function (as long as a corresponding test exists) while maintaining statistical guarantees. To this end, we

[6] Formally, given an SCM, the statement follows from the global Markov condition [Lauritzen, 1996] in the augmented graph, including an intervention node with no parents that points into X^j.

define the hypothesis

$$H_{0,\hat{f}_1,\hat{f}_2} := \{P : \mathbb{E}_{D^{sh}_{test}\sim P}\big[\mathcal{S}(D^{sh}_{test},\hat{f}_1) - \mathcal{S}(D^{sh}_{test},\hat{f}_2)\big] \leq 0, \quad p(x) = p^*(x), p(y|x) = q^*(y|x)\},$$

with the shift factor $r(x) := p^*(x)/q^*(x)$. Using data D_{test} from q^*, the methodology developed below allows us to test this hypothesis $H_{0,\hat{f}_1,\hat{f}_2}$, that is, whether, in expectation, $\hat{f}_1$ outperforms $\hat{f}_2$ in the target distribution p^*, which includes the shifted covariate distribution. In practice, the densities $p^*(x)$ or $q^*(x)$ may not be given but one can still estimate these densities from data and apply our framework using (6).

4. Testing by Resampling

In Section 3, we listed various problems that can be solved by testing a hypothesis about a shifted distribution. In this section, we outline several approaches to test a target hypothesis $\tau(Q^*) \in H_0$, see (1), using a sample $\mathbf{X}_n$ from the observed distribution Q^*. We initially consider the shift τ known, and later show that asymptotic level guarantees also apply if τ can be estimated sufficiently well from data.

Our approach relies on the existence of a hypothesis test φ_m for the hypothesis H_0 in the target domain and applies this test to a resampled version of the observed data, which mimics a sample in the target domain. We show that – under suitable assumptions – properties of the original test φ_m carry over to the overall testing procedure ψ^r_n (of combined resampling and testing, as defined in (11)).

This section is organised as follows. First, in Section 4.1, we propose a resampling scheme, which we show in Section 4.2 has asymptotic guarantees. In Section 4.3, we discuss how to sample from the scheme in practice and we describe a number of extensions in Section 4.4. In Section 4.5 we show that a simpler rejection sampling scheme can be used if stricter assumptions are satisfied.

4.1. Distinct replacement (DRPL) sampling

We consider the setting, where $\tau(q)(x) \propto r(x)q(x)$ for a known shift factor r; see (5). First, we draw a weighted resample of size m from $\mathbf{X}_n$ similar to the sampling importance resampling (SIR) scheme proposed by Rubin [1987] but using a sampling scheme DRPL ('distinct replacement') that is different from sampling with or without replacement. More precisely, we draw a resample $(X_{i_1},\dots,X_{i_m})$ from $\mathbf{X}_n$, where $(i_1,\dots,i_m) \in \{1,\dots,n\}^m$ is a sequence of distinct[7] values; the probability of drawing the sequence $(i_1,\dots,i_m)$ is

$$w_{(i_1,\dots,i_m)} \propto \begin{cases} \prod_{\ell=1}^m r(X_{i_\ell}) \propto \prod_{\ell=1}^m \frac{\tau(q)(X_{i_\ell})}{q(X_{i_\ell})} & \text{if } (i_1,\dots,i_m) \text{ is distinct and} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

[7] We use 'distinct' and 'non-distinct' only to refer to the potential repetitions that occur due to the resampling $(i_1,\dots,i_m)$ and not due to potential repetitions in the values of the original sample $\mathbf{X}_n$.

Algorithm 1 Testing a target hypothesis with known distributional shift and resampling

Input: Data $\mathbf{X}_n$, target sample size m, hypothesis test φ_m, shift factor $r(x^A)$.
1: $(i_1, \dots, i_m) \leftarrow$ sample from $\{1, \dots, n\}^m$ with weights (10) (see Appendix A.4)
2: $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n) \leftarrow (X_{i_1}, \dots, X_{i_m})$
return $\psi^r_n(\mathbf{X}_n) \coloneqq \varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n))$

We provide an efficient sampling algorithm and discuss different sampling schemes in Section 4.3. We refer to $(X_{i_1}, \dots, X_{i_m})$ as the target sample and denote it by $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)$, where U is a random variable representing the randomness of the resample. If the randomness is clear from context, we omit U and write $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)$. When m is fixed and n approaches infinity, the target sample $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)$ converges in distribution to m i.i.d. draws from the target distribution $\tau(Q^*)$; see Skare et al. [2003] for a proof for a slightly different sampling scheme. Based on our proposed resampling scheme we construct a test ψ^r_n for the target hypothesis (1) using only the observed data $\mathbf{X}_n$ by defining

$$\psi^r_n(\mathbf{X}_n) \coloneqq \varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)), \tag{11}$$

see also Algorithm 1.

We show in Section 4.2 that distinct sampling, $\Psi_{\texttt{DRPL}}$, allows us to show guarantees without introducing regularity assumptions on the test φ_m. The motivation for resampling without replacement comes from the fact that tests, as opposed to estimation of means, may be sensitive to duplicates; an extreme but instructive example is a test of the null hypothesis that no point mass is present in a distribution. A resampling test with large replacement size and possible duplicates would not be able to obtain level in such a hypothesis. Although we show in Appendix A.5 that under stricter assumptions, sampling with replacement, that is using $\Psi_{\texttt{REPL}}$, becomes asymptotically equivalent to using $\Psi_{\texttt{DRPL}}$, this example highlights, that in the non-asymptotic regime, sampling duplicates may be harmful.

The resampling scheme $\Psi_{\texttt{DRPL}}$ in (10) is similar, but not identical to what would commonly be called 'resampling without replacement' ($\Psi_{\texttt{NO-REPL}}$), where one draws a single observation X_{i_1}, removes X_{i_1} from the list of candidates for further draws and normalizes the remaining weights to reflect the absence of X_{i_1} (see also Section 4.3). $\Psi_{\texttt{DRPL}}$ and $\Psi_{\texttt{NO-REPL}}$ differ in the normalization constants, and the normalization constant in $\Psi_{\texttt{DRPL}}$ is easier to analyze theoretically. This enables Lemma A.1 in Appendix A.8, which describes the asymptotic behaviour of the mean and variance of (10) as well as of the normalization constant of (10). We consider $\Psi_{\texttt{DRPL}}$ a tool that enables simpler theoretical analysis of SIR methods; in practice it is plausible that using $\Psi_{\texttt{NO-REPL}}$ instead of $\Psi_{\texttt{DRPL}}$ will yield similar results, though we are not aware of any theory justifying this.

4.2. Pointwise asymptotic level and power

We now prove that the hypothesis test ψ^r_n inherits the pointwise asymptotic properties of the test φ in the target domain. To do so, we require two assumptions: m and n have

to approach infinity at a suitable rate, and we require the weights to be well-behaved. More precisely, we will make the following assumptions.

(A1) $m = m(n)$ satisfies $1 \leq m \leq n$, $m \to \infty$ and $m = o(\sqrt{n})$ for $n \to \infty$.

(A2) $\mathbb{E}_Q[r(X_i)^2] < \infty$.

Assumption (A1) states that m must approach infinity at a slower rate than $\sqrt{n}$. Assumption (A2) is a condition to ensure the weights are sufficiently well-behaved, and is similar to conditions required for methods based on IPW, for example [Robins et al., 2000]. If $r(x^A)$ only depends on a subset A of variables, and x^A takes finitely many values, Assumption (A2) is trivially satisfied for all Q. In the case of an off-policy hypothesis test, such as the one described in Section 3.2, a sufficient but not necessary condition for Assumption (A2) to hold for Q^* is that the policy $q^*(a|z)$ is randomized, such that there is a lower bound on the probability of each action. For a Gaussian setting, where r represents a change of a conditional $q(x^j|x^{j'})$ to a Gaussian marginal $p(x^j)$, we provide in Appendix A.9 sufficient and necessary conditions under which Assumption (A2) is satisfied. If the hypothesis of interest is in the observed domain (see Section 2.4), we are usually free to choose any target density, so we can ensure that the tails decay sufficiently fast to satisfy Assumption (A2). In Section 5.1 below, we analyze the influence of Assumptions (A1) and (A2) on our test holding level in the context of synthetic data. We now present the first main result which states that if $\alpha_\varphi := \limsup_{k\to\infty} \mathbb{P}_{P^*}(\varphi_k(\mathbf{Z}_k) = 1)$ is the asymptotic level of the test φ when applied to a sample $\mathbf{Z}_k$ from P^*, then this is also the asymptotic level of the resampling test in Algorithm 1 when applied to a sample $\mathbf{X}_n$ from Q^*. All proofs can be found in Appendix A.8.

Theorem 1 (Pointwise asymptotics – known weights)**.** *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r : \mathcal{X} \to [0, \infty)$ exists, satisfying $\tau(q)(x) = r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let φ_k be a sequence of tests for H_0 and define $\alpha_\varphi := \limsup_{k\to\infty} \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1)$. Let $m = m(n)$ be a resampling size and let ψ_n^r be the DRPL-based resampling test defined by $\psi_n^r(\mathbf{X}_n) := \varphi_m(\Psi_{DRPL}^{r,m}(\mathbf{X}_n))$, see Algorithm 1. Then, if m and Q satisfy Assumptions (A1) and (A2), respectively, it holds that*

$$\limsup_{n\to\infty} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n) = 1) = \alpha_\varphi.$$

The same statement holds when replacing both lim sup *'s with* lim inf *'s.*

Theorem 1 shows that the rejection probabilities of ψ^r and φ converge towards the same limit. In particular, the theorem states that if φ satisfies pointwise asymptotic level in the sense of (7), and Assumption (A2) holds for all $Q \in \tau^{-1}(H_0)$, then also ψ^r satisfies pointwise asymptotic level (3). Similarly, because the statement holds for $P \notin H_0$, too, ψ^r has the same asymptotic power properties as φ.

We show in Theorem 1, that Assumption (A1) is sufficient to obtain asymptotic level of the rejection procedure. In fact, as we show in the following theorem, it is also necessary: If, m, n approach infinity with $m \geq n^q$ for $q > \frac{1}{2}$, there exists a distribution

Q, a shift r and a sequence of tests φ_k such that Assumption (A2) is satisfied and $\alpha_\varphi := \limsup_{k\to\infty} \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1) < 1$ but the probability of rejecting the hypothesis on any resample of size m converges to 1. This applies for any resampling scheme, including sampling with replacement.

Theorem 2 (In general, (A1) cannot be relaxed)**.** *Fix $\ell \in \{2, 3, \ldots\}$ and let Ψ^m be any resampling scheme that outputs a (not necessarily distinct) sample of size $m = n^q$ with $q > (\ell - 1)/\ell$, and let $\alpha_\varphi \in (0, 1)$. Then there exist a distribution $Q \in \mathcal{Q}$, a distribution shift $\tau : \mathcal{Q} \to \mathcal{P}$ with a known map $r : \mathcal{X} \to [0, \infty)$, a null hypothesis $H_0 \subseteq \mathcal{P}$ and a sequence of hypothesis tests φ_k such that $\tau(Q) \in H_0$, $\limsup_{k\to\infty} \mathbb{P}_{\tau(Q)}(\varphi_k(\mathbf{Z}_k) = 1) \leq \alpha_\varphi$, $\mathbb{E}_Q[r(X)^\ell] < \infty$ and*

$$\lim_{n\to\infty} \mathbb{P}_Q(\varphi_m(\Psi^m(\mathbf{X}_n)) = 1) = 1.$$

In particular, letting $\ell = 2$ in Theorem 2 shows that Theorem 1 does not hold without Assumption (A1). On the grounds of Theorem 2, it is tempting to think that if one strengthens Assumption (A2) to $\mathbb{E}_Q[r(X_i)^\ell] < \infty$, then one can relax Assumption (A1) to $m = o(n^{1-1/\ell})$, which would enable larger resample sizes. Yet, so far, we have not succeeded in extending the current proofs of Theorem 1 to this set of assumptions. Remark A.1 after the proof of Theorem 1 provides a few details on the difficulty of extending the current proof.

So far, we have considered the case in which the known shift factor r does not depend on q. Next, we consider the setting in which the shift factor is allowed to explicitly depend on q. This is relevant, for example, if the shift τ represents a change of the conditional of a variable X^j from $q^*(x^j|x^B)$ to $p^*(x^j|x^C)$, say, but the observational conditional $q^*(x|x^B)$ is unknown, corresponding to the setting in (6). If $q^*(x^j|x^B)$ is unknown, we are not able to compute the weights $p^*(X_i^j|X_i^C)/q^*(X_i^j|X_i^B)$. However, we can still try to estimate $q^*(x^j|x^B)$ (or even r) and obtain approximate weights $\hat{r} \propto p^*/\hat{q}^*$. Assume we have two data sets $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ both containing samples from Q^*, with n_1 and n_2 observations respectively and the first one is used to estimate r and the second one to perform the test, see Algorithm A.1. Then, if we make the following modifications to Assumptions (A1) and (A2),

(A1') $m = m(n_2)$ satisfies $1 \leq m \leq n_2, m \to \infty$ and $m = o(\min(n_1^a, n_2^{1/2}))$ for $n_1, n_2 \to \infty$,

(A2') $\mathbb{E}_Q[r_q(X_i)^2] < \infty$,

the following theorem states that even when estimating the weights, it is possible to obtain pointwise asymptotic level for the target hypothesis (1) – if the weight estimation works sufficiently well.

Theorem 3 (Pointwise asymptotics – estimated weights)**.** *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift, satisfying $\tau(q)(x) \propto r_q(x)q(x)$, see (6). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let φ_k*

be a sequence of tests for H_0 *and let* $\alpha_\varphi := \limsup_{k\to\infty} \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1)$. *Let* $\hat{r}_n$ *be an estimator for* r_q *such that there exists* $a \in (0,1)$ *satisfying*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}_Q \left| \left(\frac{\hat{r}_n(x)}{r_q(x)} \right)^{n^a} - 1 \right| = 0,$$

where the expectation is taken over the randomness of $\hat{r}_n$. *Let* $m = m(n_2)$ *be a resampling size and let* $\psi^{\hat{r}}_{n_1,n_2}$ *be the DRPL-based resampling test defined by* $\psi^{\hat{r}}_{n_1,n_2} := \varphi_m(\Psi^{\hat{r}_{n_1},m}_{\texttt{DRPL}}(\mathbf{X}_{n_2}))$ *from Algorithm A.1 in Appendix A.3. Then, if* m *and* Q *satisfy Assumptions (A1') and (A2'), respectively, it holds that*

$$\limsup_{n\to\infty} \mathbb{P}_Q(\psi^{\hat{r}}_n(\mathbf{X}_n) = 1) = \alpha_\varphi.$$

The same statement holds when replacing both lim sup *'s with* lim inf *'s.*

Theorem 3 shows that $\psi^{\hat{r}}_n$ converges to the same limit as φ. In particular, as for the case of known weights, if φ satisfies pointwise asymptotic level and Assumption (A2') holds for all $Q \in \tau^{-1}(H_0)$, then also $\psi^{\hat{r}}$ satisfies pointwise asymptotic level for the hypothesis $\tau(Q^*) \in H_0$ and $\psi^{\hat{r}}$ inherits asymptotic power properties, too.

4.3. Computationally efficient resampling with $\Psi_{\texttt{DRPL}}$

In Section 4.1 we propose a sampling scheme $\Psi_{\texttt{DRPL}}$, defined by (10), and in Section 4.2 we prove theoretical level guarantees when we resample the observed data using $\Psi_{\texttt{DRPL}}$. In this section, we display a number of ways to sample from $\Psi_{\texttt{DRPL}}$ in practice.

To do so, let $\Psi_{\texttt{REPL}}$ and $\Psi_{\texttt{NO-REPL}}$ denote weighted sampling with and without replacement, respectively, both of which are implemented in most standard statistical software packages. Though $\Psi_{\texttt{DRPL}}$ and $\Psi_{\texttt{NO-REPL}}$ both sample distinct sequences $(i_1, \ldots, i_m)$, they are not equal, i.e., they distribute the weights differently between the sequences (see Appendix A.4). We can sample from $\Psi_{\texttt{DRPL}}$ by sampling from $\Psi_{\texttt{REPL}}$ and rejecting the sample until the indices $(i_1, \ldots, i_m)$ are distinct, see Appendix A.4.1. In Proposition A.1 we prove that under suitable assumptions, such as $m = o(\sqrt{n})$, the probability of drawing a distinct sequence already in a single draw approaches 1, when $n \to \infty$.

In some cases (though these typically only occur when Assumption (A1) or Assumption (A2) are violated, and our asymptotic guarantees do not apply), the above rejection sampling from $\Psi_{\texttt{REPL}}$ may take a long time to accept a sample. For these cases, we propose to use an (exact) rejection sampler based on $\Psi_{\texttt{NO-REPL}}$, which will typically be faster (since it has the same support as $\Psi_{\texttt{DRPL}}$). We provide all details in Appendix A.4.2.

If neither of the two exact sampling schemes for $\Psi_{\texttt{DRPL}}$ is computationally feasible, we provide an approximate sampling method that applies a Gibbs sampler to a sample from $\Psi_{\texttt{NO-REPL}}$; we refer to this scheme as $\Psi_{\texttt{DRPL-GIBBS}}$. Finally, one can simply approximate $\Psi_{\texttt{DRPL}}$ by a sample from $\Psi_{\texttt{NO-REPL}}$ – this is computationally faster, and leads to similar results in many cases. The details are provided in Appendix A.4.3. In practice, our implementation first attempts to sample from $\Psi_{\texttt{NO-REPL}}$ by (exact) rejection sampling,

and if the number of rejections exceed some threshold, sampling without replacement is used instead.

Proposition A.1 (mentioned above) has another implication. We prove that we can obtain the same level guarantee, when using $\Psi_{\texttt{REPL}}$ instead of $\Psi_{\texttt{DRPL}}$ (see Corollary A.1 in Appendix A.4). This result, however, requires an assumption that is stronger than Assumption (A2). Intuitively, stronger assumptions are required for $\Psi_{\texttt{REPL}}$ because sampling with replacement is much more prone to experience large variance due to observations with huge weights.

4.4. Extensions

In this section, we discuss a number of extensions of the methodology and theory presented in the preceding sections.

4.4.1. Heuristic data driven choice of m

Resampling distinct sequences requires that we choose a resampling size m that is smaller than the original sample size n. If m is too large when sampling distinct sequences, it can happen that eventually there are no more points left that are likely under the target distribution. Consequently, the resampling procedure disproportionally often has to sample points that are very unlikely in the target distribution. This leads to the target sample being a poor approximation of the target distribution. Our theoretical results show that choosing a resampling size of order $o(\sqrt{n})$ avoids this problem, see Theorem 1.

However, this result is asymptotic and does not immediately translate into finite sample statements. Furthermore, in many cases also the requirement $m = o(\sqrt{n})$ is too strict, and asymptotic level can also be obtained by setting $m = o(n^a)$ for some $a \in (1/2, 1]$ (with the most extreme case being $P^* = Q^*$, where $a = 1$ can be applied). Since a larger m typically results in increased power of the hypothesis test, we want to choose m as large as possible while maintaining that the target sample still approximates the target distribution.

Consider the case where τ corresponds to changing $q^*(x^j|x^C)$ to $p^*(x^j|x^B)$. We can then test the validity of the resampling by testing whether the target sample matches the theoretical conditional. Specifically, for a fixed m, we can verify whether the conditional $X^j \mid X^B$ in the resampled data $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)$ is close to the target conditional $p^*(x^j|x^B)$ by a goodness-of-fit test $\kappa(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)) \in \{0, 1\}$. If m is chosen too large, the resampling is likely to include many points with small weights, corresponding to small likelihoods $p^*(x^j|X^B)$, which will cause the goodness-of-fit test to reject the hypothesis that the target sample has the conditional $p^*(x^j|X^B)$.

We can use this to construct a data-driven approach to selecting m: For an increasing sequence of m's, perform the goodness-of-fit test for several resamples $\kappa(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)_1)$, …, $\kappa(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)_K)$. If $\frac{1}{K}\sum_{k=1}^{K} \kappa(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)_k)$ is smaller than some predefined cutoff `qt`, we accept m as a valid target sample size[8]. We then use the largest accepted m

[8]Concretely, since under the null hypothesis, $\kappa(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)_k)$ is uniform, we chose `qt` to be the α_c-quantile of the mean of K uniform distributions for some $\alpha_c \in (0, 1)$, see Appendix A.6. Doing so,

as the resampling size in the actual test for the hypothesis of interest. We summarize the procedure for finding m in Algorithm A.2 in Appendix A.6 and call this the GOF-heuristic for choosing m.

To avoid potential dependencies between the tuning of m and the hypothesis test, we could use sample splitting. In practice, however, we use the entire sample, since the dependence between the tuning step and the final test in our empirical analysis appears to be sufficiently low such that the level properties of the final tests were preserved, see e.g., the experiment in Section 5.1.

If the target conditional $p^*(x^j|X^B)$ is a linear Gaussian conditional density (i.e., $X^j \mid X^B \sim \mathcal{N}(\beta^\top X^B, \sigma^2)$ for some parameters β, σ under P^*) the goodness-of-fit test can be performed by using a linear regression and testing the hypothesis that the regression slope in the resample is β. For more complex conditional densities, one should prefer a test that has (pointwise asymptotic) power against a wide range of alternatives. Here, we propose to use the kernel conditional-goodness-of-fit test by Jitkrittum et al. [2020] to test that the resampled data $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n)$ has the desired conditional.

4.4.2. Combining different resamples

The proposed procedure draws a random resample and hence a different conclusion of the test may be drawn if the procedure is repeated. To reduce this randomness, we may wish to repeat the resampling and testing several times, and combine the tests into a single test.

Repeating the procedure can also have a positive impact on sample efficiency: While Theorem 1 shows that $m = o(\sqrt{n})$ suffices for any distribution shift satisfying Assumption (A2), this choice of m is in many cases too strict. An extreme case is when $P^* = Q^*$, where one could as well test the hypothesis in the n observed points instead of the m resampled points; here, using a single subsample of size less than $\sqrt{n}$ is not optimal in terms of power.

By repeated sampling, a larger part of the information contained in the data can be exploited. A difficulty in combining tests from several resamples is that the test statistics are dependent, since they are computed on random subsets of the same data set. Here, the strength of the dependence needs to be taken into account. Considering the following two corner cases may help for building intuition. If there are m weights that are much larger than the remaining ones, the draws and the test statistics of different draws are mostly identical. In this case, repeated draws do not contain additional information about the null hypothesis. If all weights are equal, however, the test statistics are less dependent and thereby the different draws contain partially complementary information.

Several procedures exist for combining dependent tests or statistics [e.g., Liu and Xie, 2020, Vovk and Wang, 2020, Rüschendorf, 1982]. Here, we propose to use the procedure proposed by Hartung [1999] that considers tests for a single hypothesis and only requires access to p-values $p_1, \ldots, p_k$ that are uniformly distributed under the null hypothesis. The procedure then transforms the p-values using the inverse Gaussian

ensures that for a fixed m, under the null hypothesis of the resample having the intended conditional, the test has level α_c.

CDF $t_i = \Phi^{-1}(p_i)$, estimates the (pairwise) covariance[9] of $t_1, \dots, t_k$ and, taking these covariances into account, considers a weighted average of the probits t_i. Under the null hypothesis, this weighted average follows a standard normal distribution, which can be used to construct a combined p-value for the null hypothesis.

Conducting repeated tests can be combined with the heuristic in Section 4.4.1: We apply the heuristic to chose a resample size m which holds level, and repeatedly test the hypothesis using this resample size, which we then combine into a single p-value. As we demonstrate empirically in Section 5.2, this often increases power compared to only testing the hypothesis once.

4.4.3. Finite-sample level guarantees

In addition to the asymptotic results presented in Section 4.2, we now prove that the hypothesis test ψ_n^r inherits finite-sample level if the test φ in the target domain satisfies finite-sample guarantees.

Theorem 4 (Finite sample level – known weights)**.** *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r : \mathcal{X} \to [0, \infty)$ exists, satisfying $\tau(q)(x) = r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let m be a resampling size and let φ_m be a test for H_0 and define $\alpha_\varphi := \mathbb{P}_P(\varphi_m(\mathbf{Z}_m) = 1)$. Also let ψ_n^r be the DRPL-based resampling test defined by $\psi_n^r(\mathbf{X}_n) := \varphi_m(\Psi_{DRPL}^{r,m}(\mathbf{X}_n))$, see Algorithm 1. Then, if Q satisfies Assumption (A2), it holds that*

$$\mathbb{P}_Q(\psi_n^r(\mathbf{X}_n) = 1) \leq \inf_{\delta \in (0,1)} \left(\frac{\alpha_\varphi}{1-\delta} + \frac{V(n,m)}{V(n,m) + \delta^2} \right), \tag{12}$$

where $V(n,m) = \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{l} \binom{n-m}{m-\ell} (\mathbb{E}_Q[r(X_1)^2]^\ell - 1)$.

Thus, if $\mathbb{E}_Q[r(X_1)^2]$ is known, one can evaluate the finite-sample level of the DRPL-based resampling test for any choice of m. We show in Appendix A.2 that the term $V(n,m)$ can be computed efficiently and such that numerical under- or overflows is avoided, even if n and m are so large that evaluating the individual terms $\binom{n}{m}$ and $\binom{m}{l}\binom{n-m}{m-\ell}(\mathbb{E}_Q[r(X_1)^2]^\ell - 1)$ may cause under- or overflows. Given $V(n,m)$, the minimization problem on the right hand side can easily be implemented in numerical optimizers or solved explicitly for the minimal δ.[10] Appendix A.2.1 contains plots of the upper bound for various values of the parameters.

[9] This assumes that the pairwise covariance $\mathrm{cov}(t_i, t_j)$ is constant for all $i \neq j$. This is satisfied in our case, since each p_i is a result of the same test.

[10] Taking the derivative with respect to δ and equating it to 0, the resulting equation can be rewritten to a polynomial equation of degree 4. One can then evaluate the right hand side of (12) at the (at most 4) roots and additionally the boundary point $\delta = 0$, and use the one that yields the smallest bound.

If τ is the identity mapping, i.e. $Q^* = P^*$, then $\mathbb{E}_Q[r(X_1)^2] = 1$, and for all m, $V(n,m) = 0$. As one would expect, in that case Theorem 4 states that for any m, $\mathbb{P}_{Q^*}(\psi_n^r(\mathbf{X}_n) = 1) \leq \alpha_\varphi$, that is, the probability of rejecting when applying φ to the resampled data is upper bounded by the probability of rejecting when applying φ directly to target data.

4.4.4. Uniform level

The asymptotic level guarantees implied by Theorem 1 are pointwise, meaning we are not guaranteed the same convergence rate for all distributions $Q \in \tau^{-1}(H_0)$. However, as the following theorem shows, if a uniform bound on the weights exists, i.e., $\sup_{Q \in \tau^{-1}(H_0)} \mathbb{E}_Q[r(X_i)^2] < \infty$, and the test φ has uniform asymptotic level, the overall procedure can be shown to hold uniform asymptotic level.

Theorem 5 (Uniform asymptotic level)**.** *Assume the same setup and assumptions as in Theorem 1. If additionally* $\sup_{Q \in \tau^{-1}(H_0)} \mathbb{E}_Q[r(X_i)^2] < \infty$ *and* $\limsup_{k \to \infty} \sup_{P \in H_0} \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1) \leq \alpha_\varphi$*, then*

$$\limsup_{n \to \infty} \sup_{Q \in \tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n) = 1) \leq \alpha_\varphi,$$

i.e., ψ_n^r *satisfies uniform asymptotic level* α_φ *for the hypothesis* $\tau(Q^*) \in H_0$*.*

4.4.5. Hypothesis testing in the observed domain

In Section 2.4, we argue that one can use our framework for testing under distributional shifts to test a hypothesis in the observed domain, too. Indeed, the results in Section 4.2 directly imply the following corollary (see Corollary A.2 in Appendix A.8.8 for a more detailed version).

Corollary 1 (Pointwise level in the observed domain)**.** *Consider hypotheses* $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$ *and* $H_0^{\mathcal{P}} \subseteq \mathcal{P}$ *and let* $\tau : \mathcal{Q} \to \mathcal{P}$ *be a distributional shift such that* $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$*. Under the same assumptions as in Theorem 1, if* φ *is a test that satisfies pointwise asymptotic level in the target domain,* ψ_n^r *satisfies pointwise asymptotic level for the hypothesis* $Q^* \in H_0^{\mathcal{Q}}$*.*

This corollary for example applies to the test in Section 3.1, where we test a hypothesis of (marginal) dependence in a shifted distribution $\tau(P)$ although the hypothesis of interest is conditional independence in Q. While the requirement in Assumption (A1) that $m = o(\sqrt{n})$ implies that we are testing the hypothesis in a smaller sample size, the hypothesis of marginal independence is also statistically simpler to test than a hypothesis of conditional independence. This also means that our approach may come with computational benefits: We can resample the data into a smaller data set, where the signal of interest is more concentrated. For example, applying a kernel conditional independence test such as the one proposed by Zhang et al. [2011] can be computationally

If $P^* \neq Q^*$ then $V(n, m) > 0$, and for any m the right hand side of (12) exceeds α_φ. To control the level of the resampling test, say at a rate α_ψ, one can set the resample size m small enough such that the right hand side of (12) is smaller than α_ψ. We propose to use the largest m such that the right hand side of (12) is bounded by α_ψ.

In practice, we find that in many settings, the inequality (12) is not strict: the largest m such that the right hand side (12) is bounded by α_ψ is not close to being the largest m such that the left hand side is bounded by α_ψ. Hence, for practical purposes, the scheme for choosing m proposed in Section 4.4.1 often returns larger values m while retaining level α_ψ under the null hypothesis; we explore this further in Section 5.

Algorithm 2 Testing a target hypothesis with known distributional shift and rejection sampling

Input: Data $\mathbf{X}_n$, hypothesis test φ_m, shift factor $r(x^A)$ and bound M.
for $i = 1, 2, \ldots, n$ **do**
  Sample U_i uniform on $(0, 1)$
  if $U_i > \frac{r(X_i)}{M}$ **then**
    Discard X_i
return $\psi_n^r(\mathbf{X}_n) \coloneqq \varphi_m(X_{i_1}, \ldots, X_{i_m})$

expensive when n is large. On the contrary, using a (marginal) kernel independence test such as the one by Gretton et al. [2008] is computationally simpler and for the proposed procedure only needs to be applied to a resample of size $o(\sqrt{n})$. In our experiments in Section 5.4, we employ both these tests and find that the procedure of resampling and testing for marginal independence is indeed orders of magnitude faster.

4.5. An alternative for uniformly bounded weights

In Section 4.1, we propose the 'distinct replacement' resampling scheme and show in Theorems 1 and 4 that this has finite and asymptotic level. The procedure requires Assumption (A2), that is that the weights have finite second moment.

We now consider the stricter assumption that the weights are globally bounded. Although this assumption is not met for most distributions that are not compactly supported, this is satisfied for example by distributions on finite state spaces. We show that, under this assumption, one can use a rejection sampler with finite sample guarantees.

Suppose that $\tau(q)(x) \propto r(x)q(x)$ and there exists a known $M \in (0, \infty)$ such that $\sup_x r(x) \leq M$. Given a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ of size n from Q^*, we can use a rejection sampler that retains observations X_i with probability $r(X_i)/M$ (and otherwise discards them) to obtain a sample from $P^* = \tau(Q^*)$, and apply a hypothesis test φ_m to the rejection sampled data; see Algorithm 2.

If φ_m has level guarantees when applied to data $\mathbf{Z}_k$ from P^*, we can test the hypothesis $\tau(Q^*) \in H_0$ with the same level guarantee, since the rejection sampled data $(X_{i_1}, \ldots, X_{i_m})$ are i.i.d. distributed with distribution P^*. We state this as a proposition.

Proposition 1 (Finite level – bounded weights)**.** *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r : \mathcal{X} \to [0, \infty)$ exists, satisfying for all x: $\tau(q)(x) \propto r(x)q(x)$ and $r(x) \leq M$. Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let φ_k be a sequence of tests for H_0 and assume there exist $\alpha_\varphi \in (0, 1)$ such that for each $k \in \mathbb{N}$: $\alpha_\varphi = \sup_k \mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1)$. Let $\psi_n^r(\mathbf{X}_n)$ be the rejection-sampling test defined in Algorithm 2. Then it holds that*

$$\mathbb{P}_Q(\psi_n^r(\mathbf{X}_n) = 1) = \alpha_\varphi.$$

4.6. Statistical inference beyond testing

In this paper, we mainly focus on statistical testing in the target domain. However, by choosing H_0 as a singleton and properly defining ϕ_k (which is not required to be a test), the result of Theorem 1 is strong enough to imply that other types of statistical inference remain valid after resampling, too. We formulate this as a corollary.

Corollary 2 (Confidence intervals, consistency, asymptotic distribution)**.** *Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r : \mathcal{X} \to [0, \infty)$ exists, satisfying $\tau(q)(x) \propto r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let g_k be an estimator or a confidence region in the target domain for a parameter $\theta(P)$. Let $m = m(n)$ be a resampling size and evaluate the estimator on the resampled data set, that is, $b_n^r(\mathbf{X}_n) := g_m(\Psi_{DRPL}^{r,m}(\mathbf{X}_n))$. Then, if m and Q satisfy Assumptions (A1) and (A2), respectively, b inherits properties like coverage, consistency or asymptotic normality from g. More precisely, we have the following three statements.*

(i) Coverage of confidence regions. Let $\alpha \in [0, 1]$ be arbitrary. Then,

$$\liminf_{k\to\infty} \mathbb{P}_P(g_k(\mathbf{Z}_k) \ni \theta(P)) \geq 1 - \alpha \quad \Rightarrow \quad \liminf_{n\to\infty} \mathbb{P}_Q(b_n^r(\mathbf{X}_n) \ni \theta(P)) \geq 1 - \alpha.$$

(ii) Consistency of estimator. Let $\varepsilon > 0$. Then, for any suitable norm $\|.\|$,

$$\lim_{k\to\infty} \mathbb{P}_P(\|g_k(\mathbf{Z}_k) - \theta(P)\| > \varepsilon) = 0 \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}_Q(\|b_n^r(\mathbf{X}_n) - \theta(P)\| > \varepsilon) = 0.$$

(iii) Asymptotic distribution of estimator. Let F be a cumulative distribution function, possibly depending on P, and let $c \in \mathbb{R}$ be such that F is continuous in x. Then,

$$\lim_{k\to\infty} \mathbb{P}_P(g_k(\mathbf{Z}_k) \leq x) = F(x) \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}_Q(b_n^r(\mathbf{X}_n) \leq x) = F(x).$$

Here, part (i) follows from Theorem 1 by choosing $\varphi_k(\mathbf{Z}_k) := \mathbb{1}_{\{g_k(\mathbf{Z}_k) \ni \theta(P)\}}$, because the proof of Theorem 1 works with any function φ_k that takes values in $\{0, 1\}$. Part (ii) follows with $\varphi_k(\mathbf{Z}_k) := \mathbb{1}_{\{\|g_k(\mathbf{Z}_k) - \theta(P)\| > \varepsilon\}}$ and part(iii) when choosing $\varphi_k(\mathbf{Z}_k) := \mathbb{1}_{\{g_k(\mathbf{Z}_k) \leq x\}}$.

4.7. Choosing resampling parameters

We summarize the choices of parameters required to apply our framework in Table 1. Given a data set of size n, we need to specify a resampling strategy and a resampling size m. Additionally, in cases where the hypothesis of interest is in the observed distribution (see Corollary A.2), we need to specify a target distribution.

If the weights are uniformly bounded, $\sup_x r(x) \leq M$ with known M, we propose to use the rejection sampler from Section 4.5, which does not require a choice of m and has finite sample guarantees (see Proposition 1). If the weights are not uniformly bounded or M is unknown, one can use the `DRPL` resampling scheme in Appendix A.4, which has asymptotic level guarantees. One can also use the `NO-REPL` resampling scheme

	Target distribution	Resampling scheme	Resampling size m
Principle	*Select p^* as close to q^* as possible in that it ensures that weights are well-behaved*	*We suggest to draw distinct observations to ensure valid inference.*	*Select m as large as possible s.t. resample remains sufficiently close to q^*.*
Details	• If target distribution p^* is not given by application, sometimes minimizing variance of weights is possible; in CI testing, often replacing $q^*(x\|z)$ by $q^*(x)$ works well • Optimize over feasible target distributions p^* to minimize $\text{Var}_{Q^*}(r(X))$	• If there exists a known M with $\sup_x r(x) \leq M$, use rejection sampling (Section 4.5) • In general, use `DREPL` • If computationally infeasible or n very large, use `NO-REPL` (assuming Assumption (A2'))	• Asymptotic theory requires $m = o(\sqrt{n})$ • Asymptotic heuristic: $m = \lfloor \sqrt{n} \rfloor$ • GOF-heuristic: m as large as possible until GOF test rejects (Section 4.4.1) (make GOF as powerful as possible to maintain overall level) • Improve power and reduce randomness by drawing multiple resamples and combining the tests (Section 4.4.2)

Table 1.: Summary of how to select the user-specified parameters of the resampling procedure.

(Appendix A.5), which is computationally faster, though sampling with replacement from a finite sample for many test statistics requires one to specify an effective sample size.

To choose the resampling size, we propose to use the *GOF-heuristic* described in Section 4.4.1, which increases m as long as a goodness-of-fit test is accepted. We propose to use conservative test by setting the rejection level high ($10\% - 30\%$), to ensure that the goodness of fit test has sufficient power to detect if the resampling did not work. One can also use the *asymptotic heuristic* $m = \lfloor \sqrt{n} \rfloor$, which ensures asymptotic level but may not be appropriate for finite samples, or the m with finite sample level guarantees from Theorem 4 which however may be too conservative in practice. We recommend to repeat the resampling and testing procedure and use the combination test from Hartung [1999] whenever possible, as described in Section 4.4.2.

In applications, where the target distributions is not given, one also needs to chose the target distribution p. For example in the application of conditional independence testing (Section 3.1) we are interested in converting a conditional distribution $q^*(x|z)$ into a marginal distribution $p(z)$. In such cases, we can choose the marginal to ensure that the weights are well-behaved, yielding a better performance of our methods. Often, using the marginal distribution in the observed distribution $p(z) = q^*(z)$ works well.

5. Experiments

We present a series of simulation experiments that support the theoretical results developed in Section 4.2 and analyze the underlying assumptions. We also apply the proposed methodology to the problems described in Section 3. A simulation experiment for model selection under covariate shift (see Section 3.6) can be found in Appendix A.7.2. Unless noted otherwise, the experiments use the $\Psi_{\texttt{DRPL}}$ resampling scheme. Code that reproduces all the experiments is available at `https://github.com/nikolajthams/testing-under-shifts`.

5.1. Exploring assumptions Assumptions (A1) and (A2)

We explore the impact of violating either Assumption (A1), stating that $m = o(\sqrt{n})$, or Assumption (A2), stating that the weights must have finite second moment in the observational distribution. To do so, we apply the procedure discussed in Section 3.1 that reduces a conditional independence test $X \perp\!\!\!\perp Y \,|\, Z$ in the observational domain to an unconditional independence test in the target domain. Specifically, we simulate $n = 10'000$ i.i.d. observations from the linear Gaussian model with

$$X := \varepsilon_X \qquad Z := X + 2\varepsilon_Z \qquad Y := \theta X + Z + \varepsilon_Y$$

for some $\theta \in \mathbb{R}$ and $\varepsilon_X, \varepsilon_Z, \varepsilon_Y \sim \mathcal{N}(0,1)$ inducing a distribution Q^* over (X, Y, Z). We assume that the conditional distribution $q^*(z|x)$ is known and replace it with an independent Gaussian distribution $\phi_\sigma(z)$ with mean zero and variance σ^2, breaking the dependence between X and Z in the target distribution.

We then perform a test for independence of X and Y in the target distribution using a Pearson correlation test. We do this both for $\theta = 0.4$ (where $X \not\perp\!\!\!\perp Y \,|\, Z$ and ideally we reject the hypothesis) and for $\theta = 0$ (where $X \perp\!\!\!\perp Y \,|\, Z$ and ideally we accept the hypothesis). Fig. 3 shows the resulting rejection rates of the test, where we have repeated the procedure of simulating, resampling and testing (at level $\alpha = 0.05$) 500 (left) or $10'000$ (right) times. In this experiment, the $\Psi_{\texttt{NO-REPL}}$ sampler is used, since the rejection samplers break as m gets very large.

First, we test the impact of changing the resampling size m. For each simulated data set $\mathbf{X}_n$ and each m, we resample 100 target data sets $\Psi^{r,m}(\mathbf{X}_n)$, and compute the rates of rejecting the hypothesis. The shaded areas in Fig. 3 (left) indicate 95% of the resulting trajectories, and the solid lines show one example simulation. Our theoretical results assume $m = o(\sqrt{n})$ and, indeed, the hypothesis rejects around 5% of simulations when $X \not\to Y$ ($\theta = 0$) for small m. As discussed in Section 4.4.1, Assumption (A1) may in some cases be too strict, and we observe that the 5% level is retained when m moderately exceeds $\sqrt{n}$; but as m grows larger, the level is eventually lost.

For the same example simulation $\mathbf{X}_n$ as in the left plot, we also apply the GOF-heuristic for choosing m as described in Section 4.4.1, and plot the resulting p-values in Fig. 3 (middle). Since the data are Gaussian, we can perform the goodness-of-fit test by a simple linear regression analysis. For each m, we compute the average of the p-values

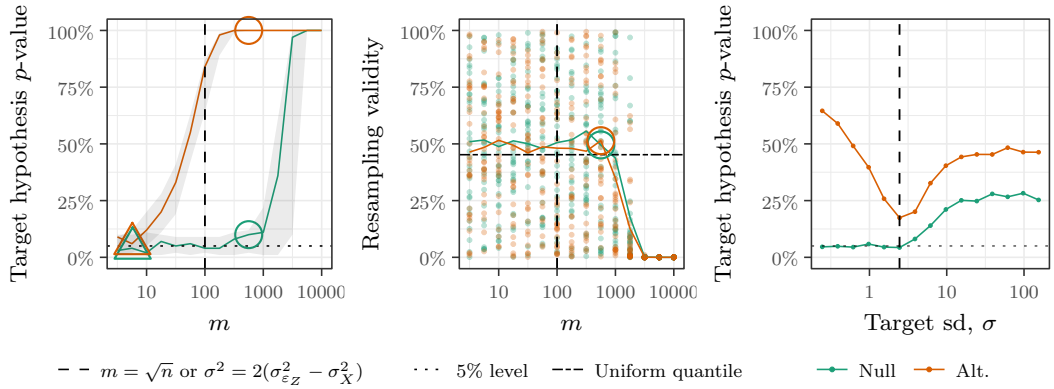


Figure 3.: (All) Rejection rates from the experiments in Section 5.1. We replace the conditional distribution $q^*(z|x)$ with a marginal distribution $\phi_\sigma(z)$. We perform a test for independence $X \perp\!\!\!\perp Y$ in the target distribution, and plot the rejection rates. (Left) Validation of Assumption (A1). We run the resampling with $\sigma = 1$ and different sample sizes m. Shaded regions show 95% quantiles of the rejection rates of the hypothesis test. The level seems to hold for $m \leq \sqrt{n}$, the latter corresponding to the asymptotic rate Assumption (A1) (left of the dashed vertical line). Circles indicate the m suggested by the middle plot and triangles the m suggested by the finite-sample method described in Section 4.4.3 – as expected, this is a conservative choice. The GOF-heuristic suggests an m that indeed yields larger power. (Middle) P-values (dots) and average p-values (lines) when applying Algorithm A.2 to choose m. We select m (circled) from the first time, the goodness-of-fit test (horizontal dashed line) is rejected. (Right) Validation of Assumption (A2). Our procedure is run with different standard deviations σ in the Gaussian target distribution $p_\sigma(z)$ $\phi_\sigma(z)$. The dashed vertical line indicates the theoretical threshold of $\sqrt{6}$, see Section 5.1.

(solid lines), and increase m until the average p-value drops below the 5% quantile of the distribution of $\texttt{mean}(U_1, \ldots, U_\ell)$ where $U_1, \ldots, U_\ell$ are i.i.d. uniform random variables. The circles in the left and middle plot indicate the m that is chosen by Algorithm 1 for this simulation. We observe in the left plot that the power of the test can be increased using the m suggested by the middle plot, while the level approximately holds at 5%.

Second, we test the importance of Assumption (A2). For different σ (and fixed $m = \sqrt{n}$), we compute the weights $r = \phi_\sigma(Z_i)/q(Z_i|X_i)$, and in Fig. 3 (right) we plot the rejection rates of the test statistic when $X \to Y$ ($\theta = 0.4$) and $X \not\to Y$ ($\theta = 0$). We show in Appendix A.9 that Assumption (A2) is satisfied if and only if $\sigma^2 < 2(\sigma^2_{\varepsilon_{2\varepsilon_Z}} - \sigma^2_X)$, where σ^2_X is the variance of X and $\sigma^2_{\varepsilon_{2\varepsilon_Z}}$ is the variance of the noise term in the structural assignment for Z. In this experiment, it follows that Assumption (A2) holds if and only if $\sigma < \sqrt{6}$. We observe that when σ exceeds the threshold of $\sqrt{6}$ (vertical dashed line), the level eventually deviates from the 5% level. Furthermore, the power drops when σ

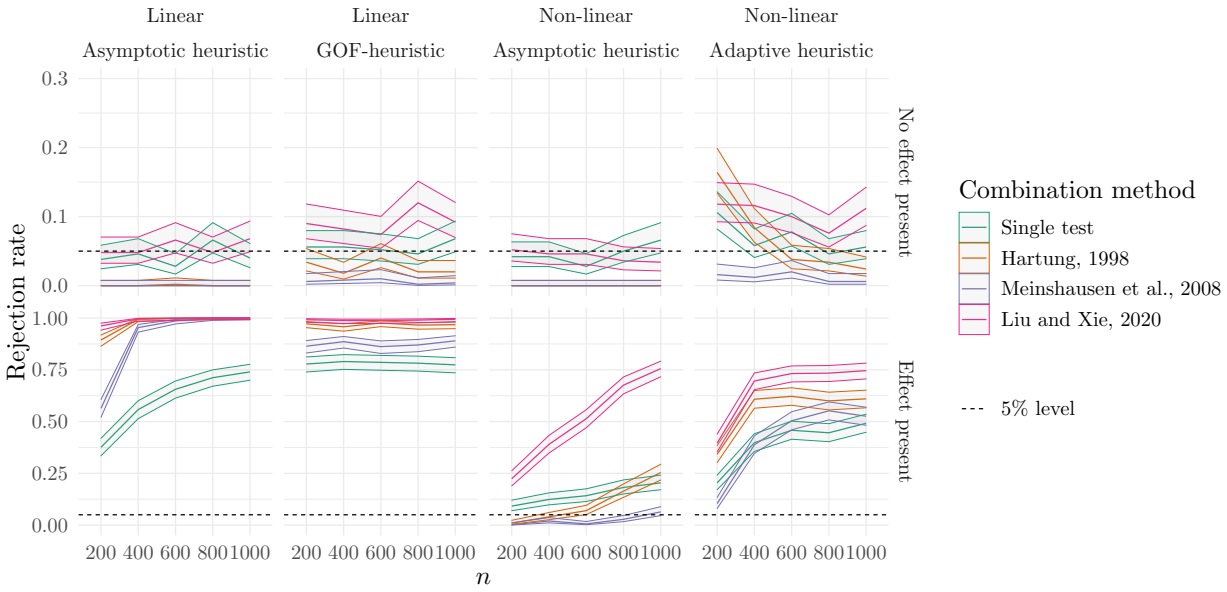


Figure 4.: Plotting rejection rates for the experiment for both a single test as well as 3 different combination tests, see Section 4.4.2.

approaches the threshold.

5.2. Combining repeated tests

In Section 4.4.2, we discussed combining multiple resamples, in order to both reduce the randomness of our resampling procedure, and to possibly get better finite sample performance, from using a larger part of the data.

In this experiment, we explore the effect of employing various combination tests, which allow for arbitrary dependencies between the tests. First, we consider the Cauchy combination tests (CCT) of Liu and Xie [2020]. They assume that the test statistics are normally distributed, though they find that the outcome is not sensitive to this assumption in practice. We also consider the combination test of Hartung [1999], which transforms the p-values (assuming they are uniform under the null hypothesis) into a Gaussian distribution, and estimates the covariance between the tests in the transformed space. Lastly, we consider the combination procedure in Meinshausen et al. [2009], which combines p-values by taking quantiles of the empirical distribution [see also Rüger, 1978, Vovk and Wang, 2020].

We consider the task in Section 3.1 of testing conditional independence. We simulate data from both a linear and a nonlinear SCM, and test the hypothesis of conditional independence $X \perp\!\!\!\perp Y \mid Z$ both in a scenario where this null hypothesis is true and where it is false. In Fig. 4, we plot the resulting rejection rates for various sample sizes n, where for each n, we sample m points, where m is selected either with the asymptotic heuristic $m = \lfloor \sqrt{n} \rfloor$ or with the GOF-heuristic (see Section 4.4.1) which increases the resample size until the resample is rejected in a goodness-of-fit test at a 10% level.

In the linear SCM, where the weights are fitted using a (correctly specified) linear model, we observe that all three combination tests hold level and gain additional power over the single test. This is a well behaved scenario, where the asymptotic heuristic is likely too conservative, and when used in conjunction with the asymptotic heuristic, the combination tests gain additional power by utilizing more information from the sample. The gain in power of using combination tests is less pronounced for the GOF-heuristic, since more information is already used by the adaptive choice of m.

In the nonlinear SCM, most types of combination tests also retain level, though in the case of the GOF-heuristic (which typically selects m more aggressively) the level is only attained for larger sample sizes or even not at all, for the test from Liu and Xie [2020]. This could be because the goodness-of-fit test, used to decide how large the resample can be, may not have sufficient power in small sample sizes to detect when a resample does not follow the intended target distribution. This nonlinear SCM, where the weights are fitted by a generalized additive model [Hastie, 2017], is less well-behaved than the linear SCM, and the weights are more degenerate; as a result, the tests are more correlated and the combination tests are not necessarily more powerful than the single test. This shows that while better power can be obtained by combining tests, it can also be lost in scenarios with degenerate weights, where the tests are strongly correlated.

5.3. Off-policy testing

We apply our method to perform statistical testing in an off-policy contextual bandit setting as discussed in Section 3.2. We generate a data set $\mathbf{X}_n$, $(n = 30'000)$, consisting of observations $X_i = (Z_i, A_i, R_i)$ with dimensions $d_Z = 3, d_A = d_R = 1$, drawn according to the following data generating process:

$$Z \coloneqq \varepsilon_Z \qquad A \mid Z \sim q^*(A|Z) \qquad R \coloneqq \beta_A^\top Z + \varepsilon_R,$$

where $\varepsilon_Z \sim \mathcal{N}(0, I_3)$ and $\varepsilon_R \sim \mathcal{N}(0, 1)$, A takes values in the action space $\{a_1, \dots, a_L\}$, where $L = 4$, $q^*(a|z)$ denotes an initial policy that was used to generate the data $\mathbf{X}_n$ and $\beta_{a_1}, \dots, \beta_{a_L}$ are parameters of the reward function corresponding to each action. A uniform random policy was used as the initial policy, i.e., for all $a \in \{a_1, \dots a_L\}$ and $z \in \mathbb{R}^3$, $q^*(a|z) = 1/L$.

The goal is to test hypotheses about the reward R if we were to deploy a target policy $p^*(a|z)$ instead of the policy $q^*(a|z)$. Here, we consider three hypotheses, namely one-sample test of means, two-sample test of difference in means and two-sample test of difference in distributions. We set the false positive rate to 5% and use $m = \sqrt{n}$ without the GOF-heuristic in all three experiments. Rejection rates are computed from 500 repeated simulations.

In the first experiment, we construct different target policies $p^*_\delta(a|z)$. For $\delta = 0$, the target policy reduces to a uniform random policy and with increasing δ, the policy puts more mass on the optimal action (and thereby increasing the deviation from the initial policy). As $\delta \to \infty$, the target policy converges to an optimal policy. More precisely, $p^*_\delta(a|z)$ is a linear softmax policy, i.e., $p^*_\delta(a|z) \propto \exp(\delta \beta_a^\top z)$. We then apply our method

to non-parametrically test whether $\mathbb{E}_{P^*_\delta}(R) \leq 0$ on the target distribution in which the policy $p^*_\delta(a|z)$ is used. For $\delta = 0$, the expected reward is zero (here, the null hypothesis is true) and for increasing δ the expected reward increases. To apply our methodology, we employ the Wilcoxon signed-rank test [Wilcoxon, 1992] in the target domain. Figure 5 (left) shows that for $\delta = 0$, our method indeed holds the correct level and eventually starts to correctly reject for increasing δ. For comparison, we include an estimate of the expected reward based on IPW.

In the second experiment, we use the same setup as in the first experiment, but now apply the two-sample testing method discussed in Section 3.3 to test whether $R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2}$, where $K = 1$ indicates a sample under the initial policy and $K = 2$ indicates a sample under a target policy. We consider two non-parametric tests, namely a kernel two-sample test based on the maximum mean discrepancy (MMD) [Gretton et al., 2012] (using the Gaussian kernel with the bandwidth chosen by the median heuristic [Sriperumbudur et al., 2009]) and the Mann-Whitney (M-W) U test [Mann and Whitney, 1947]. Here, for $\delta = 0$, the two policies coincide and for $\delta > 0$, there is a difference in the expected reward. As shown in Fig. 5 (middle), both tests are able to detect the difference. The M-W U test has more power than the MMD test.

In a third experiment, we construct different target policies $p^*_{\delta'}(a|z)$ by varying their effect on the variance of the reward distribution, while keeping the mean unchanged. More specifically, $p^*_{\delta'}(a|z)$ is a weighted random policy, i.e., $p^*_{\delta'}(a|z) \propto \delta'$ if $a = a_1$ and $\propto 1$ otherwise. This target policy yields the same expected reward as the initial policy (a uniform random policy), but yields a different variance of the reward. When $\delta' = 1$, the target policy is the same as the initial policy, whereas the variance of the reward becomes smaller when δ' increases (in Fig. 5 (right), δ' is rescaled to 0–1 range). We then apply the same two-sample testing methods used in the second experiment to test whether $R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2}$. This difference is not picked up by the M-W U test and this time, the MMD test has more power, see Fig. 5 (right).

5.4. Testing a conditional independence with a complex conditional

In the setting of conditional independence testing, we now compare our method – when turning the problem into a test for unconditional independence as discussed in Section 3.1 – to existing conditional independence tests. We sample $n = 150$ observations from the following structural causal model

$$X \coloneqq \texttt{GaussianMixture}(-2, 2) \quad Z \coloneqq -X \cdot X + \varepsilon_Z \quad Y \coloneqq \sin(Z) + \theta X^\tau + \varepsilon_Y,$$

inducing a distribution Q^*, where `GaussianMixture`(-2, 2) is an even mixture (i.e., $p = 0.5$) of two Gaussian distributions with means $\mu_1 = -2$, $\mu_2 = 2$ and unit variances, $\varepsilon_Z, \varepsilon_Y$ are independent $\mathcal{N}(0, 4)$-variables and $\theta \in [0, 3/2]$, $\tau \in \{1, 2\}$. Considering the conditional $q^*(z|x)$ to be known, we apply our methodology for testing conditional independence $X \perp\!\!\!\perp Y \,|\, Z$ with a 5% level and using the GOF-heuristic in Algorithm A.2. To do so, we replace $q^*(z|x)$ by a marginal density $\phi(z)$, which is Gaussian with mean and variance set to the empirical versions under Q^*. In the target distribution, we test

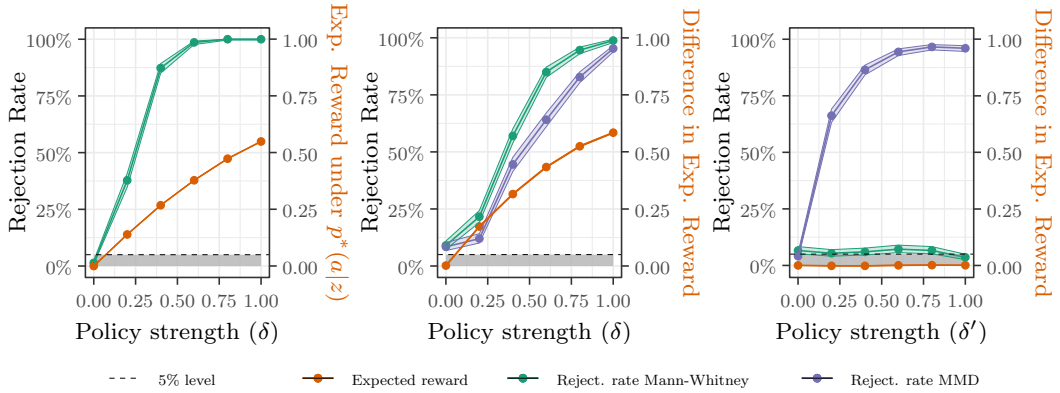


Figure 5.: Off-policy statistical testing as described in Section 5.3. (Left) One-sample test for testing whether the mean under $p^*_\delta(a|z)$ is less than or equal to 0, see Section 3.2. (Middle, Right) Two-sample tests for testing whether the reward under $p^*_\delta(a|z)$ and $p^*_{\delta'}(a|z)$, respectively, has a different distribution than the reward under the initial policy, see Section 3.3. In all cases, the null hypothesis is true for $\delta = 0$. The target policies affect the mean of the reward in the left and middle plot, whereas they affect its variance in the right plot. The framework can be combined with non-parametric tests and thereby allows for detecting complex differences in the reward distribution when comparing two policies.

for independence of X and Y using either a simple correlation test (CorTest) or a kernel independence test (HSIC) [Gretton et al., 2008]. For comparison, we also conduct conditional independence tests in the observable distribution, using the conditional permutation test (CPT) by Berrett et al. [2020], the generalized covariance measure (GCM) by Shah and Peters [2020] and a kernel conditional independence (KCI) by Zhang et al. [2011] (both using standard versions, without hyperparameter tuning). Our resampling methods use knowledge of the conditional $q^*(z|x)$, which may be seen as an unfair advantage over the conditional independence tests. Therefore, we also apply our method with estimated weights, called HSICfit, where the conditional $q^*(z|x)$ is estimated using a generalized additive model. Since CPT cannot exploit knowledge of $q^*(z|x)$, we estimate the conditional $q^*(x|z)$ to apply CPT. In this example, this is a more complex conditional than $q^*(z|x)$ [Hoyer et al., 2009, Peters et al., 2014].

We repeat the experiment 500 times and plot the rejection rates in Fig. 6 at various strengths θ of the edge $X \to Y$. All instances of our method have the correct level, see rejection rates for $\theta = 0$. When $\tau = 1$, i.e., the direct effect $X \to Y$ is linear, the power of our method approaches 100% as the causal effect increases, albeit the conditional independence tests obtain power more quickly. When the direct effect is quadratic, CorTest and GCM have little or no power, as expected since they are based on correlations (we believe that the slight deviation from 5% level in the left plot is due to very small sample sizes and the heuristic choice of m). KCI and HSIC have comparable

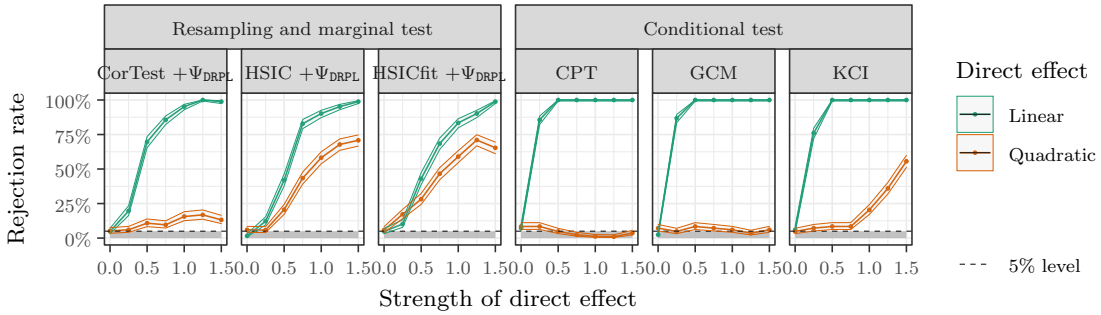


Figure 6.: Rejection rates for conditional independence tests $X \perp\!\!\!\perp Y \mid Z$ in the setting from Section 5.4. The direct effect of X on Y is θX^τ, where the exponent τ is either 1 (green) or 2 (orange), and the θ is shown on the x-axis. The three left panels show our method of resampling and testing marginal independence. Here, we combine our approach with a Pearson correlation test or HSIC. HSICfit indicates that we estimate the conditional $q^*(z|x)$ from data. We compare to three conditional independence tests, CPT, GCM and KCI. CPT seems to have a slight level violation and, as expected, GCM cannot detect the nonlinear dependence. The resampling approach can be combined with any independence test (the same holds for CPT, which is run with a correlation test, as proposed in the paper) and despite its simplicity, it is able to quickly reject the alternatives.

power in the quadratic case, with our approach even obtaining slightly more power than KCI. We observe a slight level deviation for CPT (rejecting in 8.2% of cases with a confidence interval of $(6.27\%, 11.16\%)$), which may occur because the reverse conditional $q^*(x|z)$ is harder to fit than $q^*(z|x)$. Our approach has the additional benefit of low computational costs: conditional independence testing is usually a more complicated procedure than marginal independence testing and, furthermore, the marginal test is applied to a data set of size m, which by Assumption (A1) is chosen much smaller than n.

In Appendix A.7.3 we test the same conditional independences as Berrett et al. [2020] in the bikeshare dataset by Fanaee-T and Gama [2014], and find that, when accepting hypotheses at at 5% level, our conclusions coincide with those of Berrett et al. [2020] and Candès et al. [2018].

5.5. Conditional independence testing for right-censored data

We now apply our method for conditional independence testing to the setting of right-censored data. Consider a random vector (X, Z, Y, C) with a joint density q^*, where X is a vector of covariates, Z is a vector of control variables, Y is the time to an event of interest, and C is the time to the censoring event. In the setting, we do not directly observe Y and C but instead observe the right-censored time $T = \min(Y, C)$

and the censoring indicator $\delta = \mathbb{1}_{\{T=Y\}}$. The aim is to test the conditional independence $X \perp\!\!\!\perp Y \mid Z$ given an i.i.d. sample D of (X, Z, T, δ). To the best of our knowledge, there are no non-parametric conditional independence tests available in this setting (with non-empty conditioning set). Nonetheless, there exist non-parametric marginal independence tests [e.g., Fernández et al., 2021] that we can use to test $X \perp\!\!\!\perp Y$ given the sample D. We can, therefore, apply our method for conditional independence testing introduced in Section 3.1 to conduct a non-parametric test for $X \perp\!\!\!\perp Y \mid Z$ by testing $X \perp\!\!\!\perp Y$ in the target distribution in which the conditional $q^*(x|z)$ is replaced by some marginal $\phi(x)$.

To illustrate the use of our approach, we follow Fernández et al. [2021] and consider the colon dataset as our test bed. The data are from the study of adjuvant chemotherapy in patients with stage B/C colon cancer [Laurie et al., 1989, Moertel, 1995]. Here, we consider the survival time (time to death) after the treatment as the event-time Y, which was observed for around 49% of all the patients in the study, and consider testing whether the survival time Y is independent of the covariate X (the obstruction of colon by tumour) conditioned on the set of control variables Z (age, sex and the extent of local spread).

We first analyze the dependency between the pairs $(X, Y), (X, Z)$ and (Z, Y). We use the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al., 2008] to test for independence of the pair (X, Z) and the kernel log-rank test for right-censored data [Fernández et al., 2021] to test for independence of the pairs (X, Y) and (Z, Y). Table 2a reports the p-values of the (marginal) independence tests. At the confidence level of 95%, the dependencies of all the pairs are significant. This motivates to investigate further whether the dependency between the obstruction of colon X and the survival time Y would still be significant when controlling for Z, i.e., to test for $X \perp\!\!\!\perp Y \mid Z$.

To apply our method, we estimate the conditional density $q^*(x|z)$ by $\hat{q}(x|z)$ with logistic regression and compute the weights $r = \hat{q}(x)/\hat{q}(x|z)$, where $\hat{q}(x)$ is the empirical distribution of X. The weights are then used to obtain the resample that mimics a sample from the target distribution in which the conditional $q^*(x|z)$ is replaced by $\hat{q}(x)$. The size of the resample is determined using our proposed GOF-heuristic with the threshold $\alpha_c = 0.05$ (see Algorithm A.2 in Appendix A.6). The heuristic selects $m = 675$ from the total size of $n = 929$, indicating that the two distributions are not far from each other. Table 2b reports the p-values of the independence tests in the resample. The result shows that the resampling successfully removes the dependency between X and Z in the resample (the p-value of the hypothesis $X \perp\!\!\!\perp Z$ is 0.386), while the dependency between X and Y remains significant. We can hence conclude that the dependency between the obstruction of colon X and the survival time Y is still significant when controlling for Z (age, sex and the extent of local spread).

5.6. Testing dormant independences

We now employ our method to test a dormant independence from observational data, as described in Section 3.4. We simulate data from a distribution Q^* that factorizes according to the graph $\mathcal{H}$ in Fig. 2 and test the existence of the edge $X^1 \to X^4$. As discussed by Shpitser and Pearl [2008], the presence of this edge cannot be tested by a

Hypothesis	P-value
$X \perp\!\!\!\perp Y$	0.007
$X \perp\!\!\!\perp Z$	0.028
$Z \perp\!\!\!\perp Y$	0.006

(a) Original sample

Hypothesis	P-value
$X \perp\!\!\!\perp Y$	0.002
$X \perp\!\!\!\perp Z$	0.386
$Z \perp\!\!\!\perp Y$	0.016

(b) Resample

Table 2.: p-values of independence tests on censored data in (a) the original sample and (b) in the resample that mimics a sample from the target distribution in which the conditional $q^*(x|z)$ is replaced by the marginal $\hat{q}(x)$ (corresponding to testing $X \perp\!\!\!\perp Y \mid Z$).

conditional independence test, and instead we test marginal independence between X^1 and X^4 in the target distribution $Q^{\mathrm{do}(X^3:=N)}$, which can be obtained by applying our method using $r_q(x^3, x^2) := q^{\mathrm{do}\,(X^3:=N)}(x^3)/q(x^3|x^2)$, where $q^{\mathrm{do}(X^3:=N)}(x^3)$ is the density in the intervention distribution.

More precisely, we conduct three experiments. In the first experiment, we consider binary random variables for the observables X^1, X^2, X^3 and X^4, while the hidden variable H is a discrete random variable with 4 possible values. We estimate $q(x^3|x^2)$ by the empirical probabilities and use the empirical marginal distribution of X^3 as a target distribution, i.e., $p(x^3) = \hat{q}(x^3)$. In general, choosing the marginal distribution as a target distribution corresponds to a relatively small empirical variance of the weights $r_q(x^3, x^2)$, see Assumption (A2), even though it may not always correspond to the minimum.[11] We employ Fisher's exact test to determine whether X^1 and X^4 in the interventional distribution are independent of each other. We compare our method to a more specialized method based on binary nested Markov models [Shpitser et al., 2012] that is based on a likelihood ratio test. Appendix A.7.4 contains simulation parameters from all three experiments.

In the second experiment, we consider a linear Gaussian SCM. We estimate the conditional $q(x^3|x^2)$ by a linear regression and, as before, use the empirical marginal distribution of X^3 as a target distribution. We then test for independence between X^1 and X^4 in the interventional distribution using a simple correlation test. Since the distribution is jointly Gaussian (with linear functions) and satisfies the 'bow-free' condition [Brito and Pearl, 2002b], there is a specialized, non-trivial likelihood procedure for model selection that we can compare with: We perform maximum likelihood estimation as suggested by Drton et al. [2009] and use the penalty from Nowzohour et al. [2017] to score the graphs $\mathcal{G}$ and $\mathcal{H}$.

In the third experiment, we consider a nonlinear SCM with non-Gaussian errors. We estimate the conditional $q(x^3|x^2)$ using generalized additive models. For simplicity, we consider a distribution where $X^3 - \mathbb{E}[X^3 \mid X^2]$ is Gaussian. Our procedure also applies

[11] In the special case of binary variables, when the density $q(x^3, x^2)$ is known, one can compute the marginal which minimizes the variance: it outputs one with probability $\frac{q(x^2=1)^2/q(x^3=0,x^2=1)+q(x^2=0)^2/q(x^3=0,x^2=0)}{q(x^2=1)^2/q(x^3=1,x^2=1)+q(x^2=0)^2/q(x^3=1,x^2=0)+q(x^2=1)^2/q(x^3=0,x^2=1)+q(x^2=0)^2/q(x^3=0,x^2=0)}$.

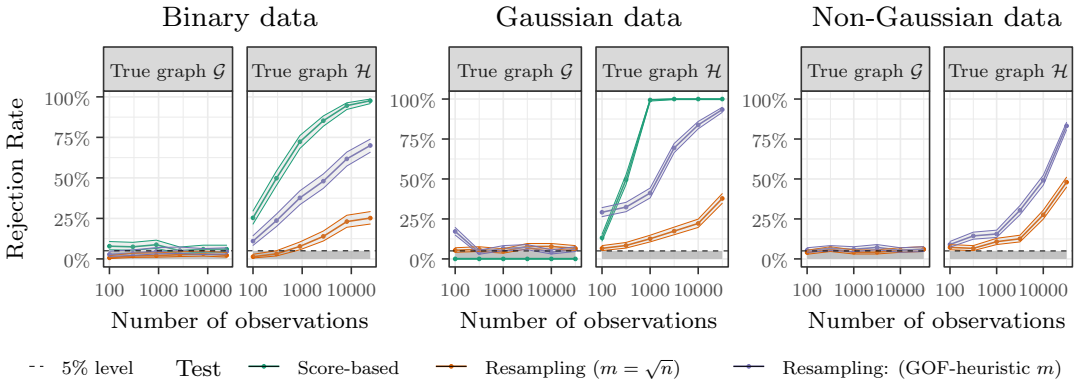


Figure 7.: We test the dormant independence discussed in Section 3.4. The existence of the edge $X^1 \to X^4$ can be inferred by testing a Verma constraint in the observational distribution, which translates to an independence statement in the target distribution. The plots show rejection rates for the underlying graphs $\mathcal{G}$ (first, third and fifth plot) and $\mathcal{H}$ (second, fourth and sixth plot), see Fig. 2, as the number of observations increases. Tailored score-based approaches exist in the binary and Gaussian cases (left and middle) but not in the case of more complex distributions (right). We plot the rates both when $m = \sqrt{n}$ and when m is chosen according to the GOF-heuristic in Algorithm A.2. The GOF-heuristic chooses an m that yields a good balance between level (graph $\mathcal{G}$) and power (graph $\mathcal{H}$).

to more general settings by applying conditional density estimation to learn $q(x^3|x^2)$, for example. To the best of our knowledge, there exist no other methods for testing the dormant independence in any of such cases.

In all experiments, we consider two strategies for choosing the resampling size m: (1) $m = \sqrt{n}$ and (2) the GOF-heuristic (see Algorithm A.2). The resulting rejection rates over 500 repeated experiments, for several sample sizes, are shown in Fig. 7. Our method identifies both the absence and presence of the causal edge $X^1 \to X^4$ in both the binary and the Gaussian setting. In both the binary and Gaussian settings, the tailored score-based approaches have more power to detect the absence of the edge (though in the binary case, the level of the test does not seem to hold exactly when the sample sizes are small). For the general case (nonlinear and non-Gaussian), our method has the correct level and increasing power as sample size increases. We are not aware of any other existing test that can achieve this in general. Compared to $m = \sqrt{n}$, the choice of m with the GOF-heuristic yields larger test power without sacrificing too much the level of the test (although the level is violated for small sample sizes in the Gaussian setting).

5.7. Testing front-door assumptions

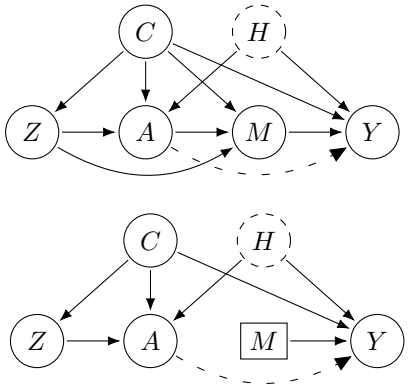
We apply our method to testing conditional independence under interventions to test the front-door assumptions as proposed by Bhattacharya and Nabi [2022] on the Framingham heart study dataset [Dawber et al., 1951]. We revisit this dataset and apply our proposed procedure for choosing the resample size m (see Section 4.4.1) and the method for combining different resamples (see Section 4.4.2). The Framingham heart study is a longitudinal epidemiology study of the risk factors for cardiovascular disease. Here, we consider a similar setup as Bhattacharya and Nabi [2022] and consider testing the front-door assumptions for estimating the effect of smoking (treatment A) on the development of coronary heart disease (outcome Y). The hypertension condition is considered as a mediator M and the past history of hypertension as an anchor Z. In addition, the set of covariates C containing age, sex, and BMI are included as control variables. Fig. 8 (top left) illustrates the assumed underlying causal graph. To apply the front-door adjustment for estimating the causal effect of A on Y, one major assumption is that there is no direct edge from A to Y. Under faithfulness [Pearl, 2009], this assumption can be verified by testing $Z \perp\!\!\!\perp Y \mid C$ in the interventional distribution in which M is replaced by some marginal, see Fig. 8 (bottom right). We apply our resampling approach by first estimating the conditional $\hat{q}(m|a,z,c) \approx q^*(m|a,z,c)$ with logistic regression and computing the weights $r = \hat{q}(m)/\hat{q}(m|a,z,c)$, where $\hat{q}(m)$ is the empirical distribution of M. The resample size m is chosen using the GOF-heuristic with $\alpha_c = 0.05$ (see Algorithm A.2). In addition, we compute p-values from multiple resamples and combine them as discussed in Section 4.4.2.

Figure 8 (right) presents p-values of the (conditional) independence tests in the original sample (a) and in the resample (b). We use the kernel conditional independence (KCI) test [Zhang et al., 2011] and the Hilbert Schmidt independence criterion (HSIC) test [Gretton et al., 2008] for the hypotheses $Z \perp\!\!\!\perp Y \mid C$ and $M \perp\!\!\!\perp (A, C, Z)$, respectively. The result suggests that the front-door assumption is plausible (the p-value of the hypothesis $Z \perp\!\!\!\perp Y \mid C$ is 0.263). (Instead of a conditional independence test under intervention, Bhattacharya and Nabi [2022] use a slightly more involved testing procedure and come to the same conclusion.) Moreover, our approach successfully eliminates the effect from (Z, A, C) on M in the resample (the p-value of the hypothesis $M \perp\!\!\!\perp (A, C, Z)$ is 0.295).

5.8. Comparison to IPW

Inverse probability weighting (IPW) allows us to test simple hypotheses such as $\mathbb{E}_P[f(X)] = c$ for some constant $c \in \mathbb{R}$ and a given function f. If data $\mathbf{Z}_m$ sampled from the target distribution P^* are available, we could test the hypothesis using the test statistic $\frac{1}{m}\sum_{i=1}^{m} f(Z_i)$. If, instead, data are available from a different observable distribution Q^*, we can estimate the corresponding test statistic in the target domain using the test statistic

$$T(\mathbf{X}_n) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \bar{r}(X_i) f(X_i),$$



Hypothesis	P-value
$Z \perp\!\!\!\perp Y \mid C$	0.001
$M \perp\!\!\!\perp (A, C, Z)$	0.001

Hypothesis	P-value
$Z \perp\!\!\!\perp Y \mid C$	0.263
$M \perp\!\!\!\perp (A, C, Z)$	0.295

Figure 8.: (Top left) The assumed causal graph in the observational distribution. (Bottom left) The assumed causal graph in the target distribution, where incoming edges in M has been resampled away. (Top right) p-values of the (conditional) independence tests in the observed distribution. (Bottom right) p-values of the (conditional) independence tests in the target distribution.

where $\bar{r}(X_i)$ is the normalized versions of the shift factor r (elsewhere we do not require r to be normalized). If $r(X_i)f(X_i)$ has finite second moment in Q, then $T(\mathbf{X}_n)$ is asymptotically normal with mean $\mathbb{E}_P[f(X_1)]$ and variance $\sigma^2 := \mathbb{V}_Q(r(X_1)f(X_1))/\sqrt{n}$, and one can construct a $(1-\alpha)$ confidence interval as

$$[T(\mathbf{X}_n) - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \quad T(\mathbf{X}_n) + z_{\alpha/2} \cdot \sigma/\sqrt{n}],$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile from the standard normal distribution.

To compare our approach to the IPW approach, we simulate data ($n = 100$) from the following structural equation model

$$X^1 := 1 + \varepsilon_{X^1} \quad X^2 := X^1 + \varepsilon_{X^2} \quad X^3 := X^2 - X^1 + \varepsilon_{X^3},$$

with $\varepsilon_{X^1} \sim \mathcal{N}(0, 3)$, $\varepsilon_{X^2} \sim \mathcal{N}(0, 4)$ and $\varepsilon_{X^3} \sim \mathcal{N}(0, 1)$. In this model, the mean of X^3 is $\mathbb{E}_Q[X^3] = 0$. We consider the distributional shift corresponding to the intervention $\text{do}(X^2 := \mu + \bar{\varepsilon}_{X^2})$ with $\bar{\varepsilon}_{X^2} \sim \mathcal{N}(0, 1)$, where X^3 has mean $\mathbb{E}_P[X^3] = \mu - 1$, and test the hypothesis $\mathbb{E}_P[X^3] = 0$ for various μ using both our resampling approach (with m chosen according to the GOF-heuristic in Algorithm A.2) and the IPW based confidence intervals. Since IPW is sensitive to degenerate weights, we also use a 'clipped IPW', where we truncate the 10 largest weights at the 10th largest value (see e.g. Cole and Hernán [2008]).

Ideally, we accept the hypothesis for $\mu = 1$ and reject the hypothesis for all other μ. The larger μ becomes, the easier it should be to reject the hypothesis $\mu = 1$, if target data are available. At the same time, since the target distribution is a Gaussian distribution centered at $\mu - 1$, as μ increases, the weights get increasingly degenerate, because the weights of the data points with the largest numerical values X^2 dominate

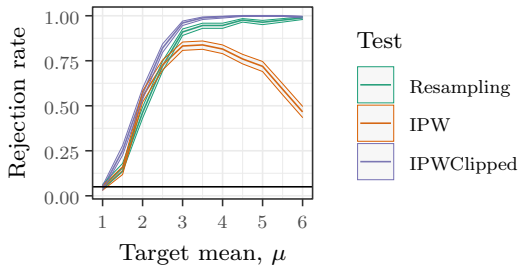


Figure 9.: Rejection rates for testing whether the mean of X^3 equals zero, which is the case if and only if $\mu = 1$, see Section 5.8. As μ increases, the weights become more and more ill-behaved and IPW loses some of its power. Neither the clipped version of IPW nor the resampling framework suffer from this problem.

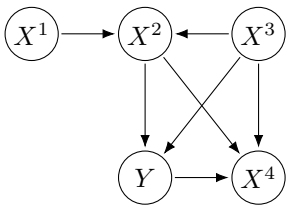
the weights of all other data points.

We observe in Fig. 9 that all methods have the correct level at 5% (when $\mu = 1$) and approximately the same power for small μ. As μ grows, the plain IPW loses its power, due to weight degeneracy. Both the clipped IPW and our resampling approach do not suffer from this issue, with power approaching 1, even as weights get increasingly degenerate. This experiment indicates that our method may share some of the robustness to degenerate weights that is known from clipped IPW, and at the same time is able to estimate more complex test statistics, that cannot be estimated using IPW.

5.9. Resampling for heterogeneity to identify causal predictors

In Section 3.5, we propose to use our resampling approach to create heterogeneous data We therefore generate $n = 1'000$ i.i.d. observations of Y, X^1, X^2, X^3, X^4 according to a linear Gaussian SCM with the graphical representation given in Fig. 10 (left). Furthermore, we assume that the conditional distribution $q^*(x^2|x^1, x^3)$ is known (instead, one could also assume that $\mathrm{PA}_2 = \{1, 3\}$ is known and estimate the conditional). As described in Section 3.5, we now generate two environments by considering the observational distribution and a modified distribution based on a distributional shift. Specifically, we take the entire sample to form environment $K = 1$ and then, to form environment $K = 2$, we resample from the same data $m = 30$ (approximately $\sqrt{n}$) observations under the distributional shift generated by replacing the conditional $q^*(x^2|x^1, x^3)$ with the target distribution $p^*(x^2|x^1, x^3)$ (which flips the sign of the dependence on x^3). The precise data generating process is described in Appendix A.7.4. This results in a data set with $n + m$ observations from two environments. We then apply ICP to the joint data from both environments and output the following estimate of the causal predictors:

$$\hat{S} := \bigcap_{S: H_{0,S} \text{ accepted}} S,$$



Set	Frequency
$\emptyset$	15.4%(11.6%)
$\{1,2,3,4\}$	0.2%(0.0%)
$\{2\}$	54.2%(40.6%)
$\{2,3\}$	26.2%(42.0%)
$\{2,3,4\}$	0.0%(1.6%)
$\{2,4\}$	0.2%(0.0%)
$\{3\}$	3.8%(4.2%)

Figure 10.: (Left) Graphical representation of the SCM used in the simulation of Section 5.9. We assume that the conditional $q^*(x^2|x^1, x^3)$ is known and use this to generate resamples that mimic a heterogeneity in the data that we can then exploit for causal discovery. (Right) Frequencies of the estimated sets. The theory guarantees that in at most 5% of the cases, the estimated set is not a subset of $\{2, 3\}$, the correct set of causal predictors. The numbers in parenthesis are oracle benchmarks that sample the data directly from the actual target distribution instead of using resampling.

Here, $H_{0,S}$ is the hypothesis defined in Section 3.5. We use the `InvariantCausalPrediction` R-package for this experiment, which tests $H_{0,S}$ using a Chow test [Chow, 1960]. We repeat the experiment 500 times and report in Fig. 10 (right) how many times each set S is output As an oracle benchmark, we also report the corresponding frequencies when we sample the target distribution directly, instead of resampling it (in particular we use the same total sample size $m + n$). Our method frequently returns the invariant set $\{2, 3\}$ and holds the predicted coverage guarantee: in only 4.2% of the cases, the estimated set is not a subset of $\{2, 3\}$.

The output of the method is guaranteed to be (with large probability) a subset of the set of true causal predictors, but depending on the type of heterogeneity, the method may output the empty set. E.g., if the true (unknown) underlying graph equals $X^4 \to Y \to X^1 \to X^2 \leftarrow X^3$, then (for the same experiment), both ICP based on the resampled data and ICP based on the true target distribution always output the empty set.

The difference between the oracle method and the resampling method (see Fig. 10) indicates that the resampled distribution does not equal the target distribution. Indeed, in some regions where the target density has substantial mass, there are no data points that can be sampled. This, however, does not show any effect on the level of the overall procedure. Thus, in the resampled data the conditional distribution of X^2, given X^1 and X^3 differs from $q^*(x^2|x^1, x^3)$ (even though it does not equal the target conditional $q^*(x^2|x^1, x^3)$). We hypothesize that the result is therefore similar to choosing a different target distribution in the first place. Indeed, when changing $p^*(x^1, x^2, x^3)$ to match the data support, the set frequencies of the oracle version closely match the resampled version.

5.10. Comparing to double machine learning methods in treatment effect estimation

We consider a treatment effect estimation setup, where for a binary treatment D, observed confounders X and continuous outcome Y we have that

$$\mathbb{P}_{Q^*}(D = 1) = m_0(X) \qquad Y = g_0(D, X) + \varepsilon_Y \qquad \mathbb{E}_{Q^*}[\varepsilon_Y | X, D] = 0$$

for some unknown functions m_0 and g_0. We consider a setting, where X is 20-dimensional, m_0 is a sigmoid function and g_0 is either a linear function or a complex, nonlinear function.

We then consider the hypothesis (in the observed distribution) that the average treatment effect (ATE) is zero, where the ATE is given by $\theta_0 = \mathbb{E}_{Q^*} g_0(1, X)] - \mathbb{E}_{Q^*}[g_0(0, X)]$. Semi-parametrically efficient estimation of θ_0, using doubly robust methods that model both m_0 and g_0 is possible if estimators of these converge sufficiently fast [Robins et al., 1994, Robins and Rotnitzky, 1995, Chernozhukov et al., 2018].

In this experiment, we compare our resampling methodology to a doubly robust method, where for the latter we fit a logistic regression model to estimate m_0 and either a random forest (nonlinear case) or a linear regression (linear case) to estimate g_0, using the `DoubleML` package in Python [Bach et al., 2022]. We then test the hypothesis $\theta_0 = 0$ by checking whether a 95% confidence interval contains zero.

To apply our methodology to this problem, we also fit a logistic regression model to estimate the conditional $q^*(d|x)$ and replace this conditional with a marginal target probability $p(d)$, which matches the empirical marginal probability of treatment, $q^*(d)$. We then apply a t-test φ_m to the resample [Student, 1908], to test the hypothesis of $\theta_0 = 0$ at a 5% level using the combination test from Hartung [1999], as discussed in Section 4.4.2. We apply this method both with the asymptotic heuristic ($m = \lfloor \sqrt{n} \rfloor$) and with the GOF-heuristic, where we chose m as large as possible while still accepting a goodness-of-fit test that the resample follows the target distribution (Section 4.4.1) at a 20% level.

We repeat the experiment 300 times and in Fig. 11 we plot the rejection rates for the hypothesis both in a scenario where $\theta_0 = 0$ (no effect present) and where the ATE is non-zero (effect present). Our procedure, using either heuristic, and the double ML procedure satisfy the desired 5% level for all sample sizes. When the true outcome model g_0 is a complex function, our procedure with the GOF-heuristic outperforms the power of the double ML method; our resampling procedure does not model g_0, and so is unaffected by how difficult it is to estimate g_0. On the contrary, when g_0 and m_0 are both simple functions, the double ML approach (which we correctly specify by using a linear and a logistic regression) attains more power than our procedure. The GOF-heuristic has more power than the asymptotic heuristic in both the linear and the nonlinear setting, as one would expect since the GOF-heuristic use larger resampling sizes m to test the hypothesis.

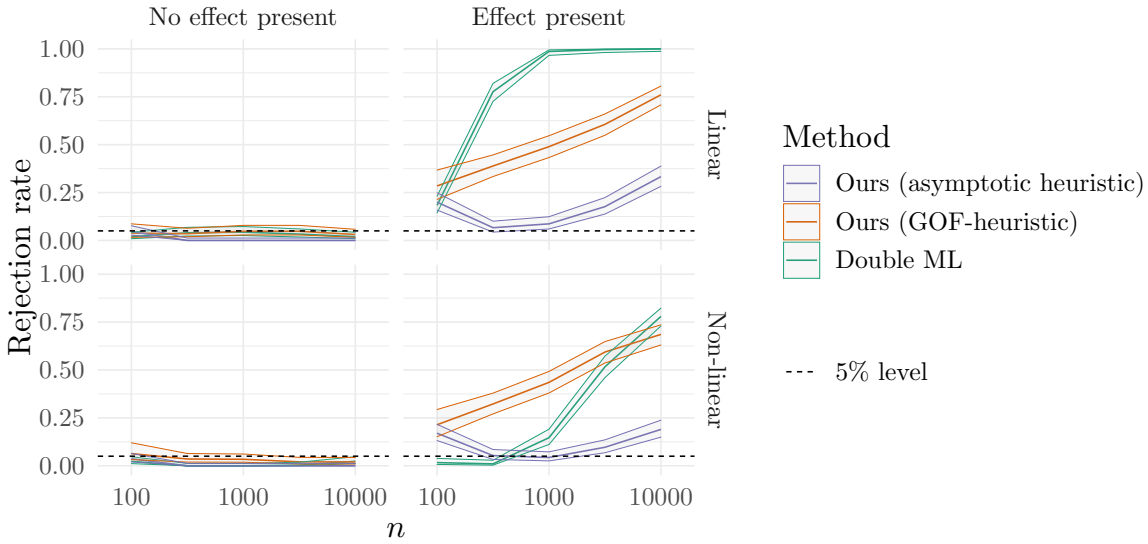


Figure 11.: Rejection rates for the hypothesis that the average treatment effect (ATE) is 0 in the experiment in Section 5.10. We consider two different data generating mechanisms (rows) and both a situation where the treatment has no effect and one where it does (columns).

6. Conclusion and Future Work

We formally introduce statistical testing under distributional shifts and illustrate that it can be applied in a diverse set of areas such as contextual bandits, conditional independence testing and causal inference. We provide a general testing procedure based on weighted resampling and prove pointwise asymptotic level guarantees under mild assumptions. Our simulation experiments underline the usefulness of our method: It is able to test complicated hypotheses, such as dormant independences – for which to-date no test with provable level guarantees exists – and can be applied to test complex hypotheses in off-policy testing or covariate shift. The framework is competitive even in some of the problems, where more specialized solutions exist. Its key strength is that it is very easy to apply and can be combined with any existing test making it an attractive go-to method for complicated testing problems.

We believe that several directions would be worthwhile to investigate further. In many of the empirical experiments, the requirement that $m = o(\sqrt{n})$ seems too strict and can be relaxed, see also Section 4.4.1. As discussed in Section 4.2, we hypothesize that under further restrictions on the weights or the test statistics, the assumption for the theoretical results can be relaxed. Bickel et al. [2012] consider the '$\binom{n}{m}$' bootstrap, which resamples distinct sequences without weights, and show that under mild assumptions, bootstrap estimates converge if $m = o(n)$. Further work is required to extend this to the case of weighted samples.

We show in Section 4.4.4 that the main convergence result, Theorem 1, can be ex-

tended to uniform level, if we make uniform assumptions on the target test, φ, and that the weights are uniformly bounded over $\tau^{-1}(H_0)$. In many model classes, the latter assumption may be too strict, and a better understanding of necessary conditions would help.

While Theorem 3 provides guarantees when $r(x^A)$ is unknown, the theorem requires guarantees on the relative error $\hat{r}(x^A)/r(x^A)$. A more natural guarantee would be on the absolute error $|\hat{r}(x^A) - r(x^A)|$, and we hope further work can shed light on the appropriate conditions (such as model classes p and q or properties of the estimator $\hat{r}$) to achieve such a guarantee.

Resampling distinct sequences is less prone to weight degeneracy than IPW or resampling with replacement, but in setups with well-behaved weights this may come at a cost of power when resampling only $m \ll n$ points. Resampling non-distinct sequences share many similarities with IPW (for fixed n, expectations of $\Psi^{r,m}_{\texttt{REPL}}$ converge to the IPW estimate when $m \to \infty$), but additionally benefits from the ability to test hypotheses where the test statistic cannot be written as an average over the data points (see Section 5.8). Further investigation of the differences between the sampling schemes and benefits and disadvantages in comparison to IPW is needed.

Our methodology considers the setting where we only observe data $\mathbf{X}_n$ from the distribution Q^*. If additionally a sample $\mathbf{Z}_{n'}$ from the target distribution P^* is already available, one can combine the two data sets, to get a larger approximate sample from P^*, a problem known as 'domain adaptation' in the literature [Finn et al., 2017]. In particular, if $\mathbf{Z}_{n'}$ is also available, one could perform the testing on the combined data set $(\Psi(\mathbf{X}_n), \mathbf{Z}_{n'})$. We believe that similar theoretical guarantees can be proved.

When testing for a hypothesis in the observed domain, we often have the freedom to choose a target distribution which could help us improve the performance of our test (as discussed in Section 2.4). In the experiments, e.g., Section 5.3 and Section 5.4, we choose the target distribution that matches certain marginals, which often helps to minimize the variance of the weights. Another possibility is to choose a target distribution such that the alternative becomes easier to detect which can be achieved by minimizing the p-value of the test with respect to the choice of the target distribution.

Acknowledgments

We thank Peter Rasmussen for valuable ideas about the combinatorics, Mathias Drton for helpful discussions on hidden variable models, Tom Berrett for insightful comments during a discussion of an earlier version of this paper and Rohit Bhattacharya for providing us the code for preprocessing the Framingham heart study data. NT, SS, and JP were supported by a research grant (18968) from VILLUM FONDEN and JP was, in addition, supported by the Carlsberg Foundation. NP was supported by a research grant (0069071) from Novo Nordisk Fonden.

3. Shifts in Distributions: Prediction

This chapter contains the following three papers:

[ShiftEval] [Thams et al., 2022a]. N. Thams, M. Oberst, and D. Sontag. Evaluating robustness to dataset shift via parametric robustness sets. In *Neural Information Processing Systems (NeurIPS)*, 2022a. NT and MO contributed equally, order determined by coin flip.

[ProxyAR] [Oberst et al., 2021]. M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

[InvPolicy] [Saengkyongam et al., 2021]. S. Saengkyongam, N. Thams, J. Peters, and N. Pfister. Invariant policy learning: A causal perspective. *arXiv preprint arXiv:2106.00808*, 2021.

[ProxyAR] and **[InvPolicy]** are concerned with learning models for predicting a target Y from covariates X, in such a way that one not only optimizes for model performance in the training distribution Q, but also in a test distribution P. Tuning models to out-of-distribution performance requires a trade-off between on one hand good predictive power in Q and on the other hand not depending on spurious correlations in Q that are not present elsewhere.

In-distribution performance in Q can be optimized by minimizing the squared error on the training data, to learn the conditional mean $\mathbb{E}_Q[Y|X]$. By definition, this predictor would achieve the smallest possible squared error on the test data if that test data were also generated from Q. However if $P \neq Q$, the conditional mean $\mathbb{E}_Q[Y|X]$ may be utilizing dependencies in Q that do not transfer to other distributions. Assume for example that the training data is generated by the SCM $X_1 \to Y \to X_2$ and that P differs from Q by an intervention which shifts the mean of both X_1 and X_2. While X_2 is predictive for Y in Q, the dependence between Y and X_2 will not be the same in P as it is in Q, and so $\mathbb{E}_Q[Y|X_1, X_2]$ may not be a good predictor of Y in P. On the contrary, because the conditional $Y|X_1$ is the same in Q and P (see (3)), the conditional mean $\mathbb{E}_Q[Y|X_1]$ remains unchanged even if the mean of X_1 shifts between Q and P; this motivates the use of invariant models where the conditional mean $\mathbb{E}[Y|X]$ remain fixed across environments [Peters et al., 2016, Magliacane et al., 2018].

In **[InvPolicy]**, we consider a contextual bandit setting and use exogenous environment variables e to represent shifting environments. We explore 'how different' P and Q

have to be, in order for an invariant model to outperform non-invariant models. On one hand, P and Q cannot be too different: Assumption 3 requires that invariant conditioning sets in the training environments are also invariant in the test environments. This is similar to the assumption that causal effects are stable across environments, discussed in the beginning of Chapter 1. On the other hand, P and Q cannot be too identical either: Assumption 2 ensures that there exist test environments where the dependence between predictors and confounders has changed. We show that if assumptions are met, then invariant models are minimax optimal over the test environments, i.e. the expected reward in the worst-case environment is larger when using an optimal invariant policy than when using an optimal non-invariant policy.

Instead of studying whether or not invariant models are minimax optimal, in **[ProxyAR]** we enforce *some* amount of invariance by minimizing a combination of the squared error in Q and a measure of invariance, following the same approach as Rothenhäusler et al. [2021]. One can show that the resulting predictor is minimax optimal over the class of distributions that arise due to mean shifts in an exogenous variable A. We examine the impact of not directly measuring A, but only measuring a proxy W of A: We show in Theorem 1 that this implies that the estimator is minimax optimal over a smaller collection of distributions than if A had been observed, where the reduction factor is the signal-to-variance ratio in W. Further we show that if we measure not just one, but two proxies W and Z of A (and provided that they are conditionally independent $W \perp\!\!\!\perp Z|A$) then we can recover the same minimax guarantee as if A had in fact been observed.

[ShiftEval] does not aim to learn a minimax optimal model at all. Instead, for a given model, it estimates how a distribution shift impacts the performance of the model. We propose a parameterization of distribution shifts by using exponential family models. We aim at finding the worst-case shift within some bounded distance from the observed distribution, and discuss how this in principle could be solved by importance sampling (introduced in Section 1.4 in Chapter 1), although non-convexity may make this optimization problem infeasible and potentially large variance of importance sampling may make it imprecise. Instead of importance sampling, we use a second order Taylor expansion to approximate the loss in the shifted distribution. Because of the assumption of exponential families, the Taylor approximation can be computed as covariances in the data. The resulting worst-case optimization problem is a quadratic program and although this problem is also non-convex, it is generally computationally feasible.

Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

Nikolaj Thams*, Michael Oberst* and David Sontag

Abstract

We give a method for proactively identifying small, plausible shifts in distribution which lead to large differences in model performance. To ensure that these shifts are plausible, we parameterize them in terms of interpretable changes in causal mechanisms of observed variables. This defines a parametric robustness set of plausible distributions and a corresponding worst-case loss. While the loss under an individual parametric shift can be estimated via reweighting techniques such as importance sampling, the resulting worst-case optimization problem is non-convex, and the estimate may suffer from large variance. For small shifts, however, we can construct a local second-order approximation to the loss under shift and cast the problem of finding a worst-case shift as a particular non-convex quadratic optimization problem, for which efficient algorithms are available. We demonstrate that this second-order approximation can be estimated directly for shifts in conditional exponential family models, and we bound the approximation error. We apply our approach to a computer vision task (classifying gender from images), revealing sensitivity to shifts in non-causal attributes.

1. Introduction

Predictive models may perform poorly outside of the training distribution, a problem broadly known as dataset shift [Quionero-Candela et al., 2009]. In high-stakes applications, such as healthcare, it is important to understand the limitations of a model in advance [Finlayson et al., 2021]: given a model trained on data from one hospital, how will it perform under changes in the population of patients, in the incidence of disease, or in the treatment policy?

In this paper, our goal is to **proactively** understand the sensitivity of a predictive model to dataset shift, using only data from the training distribution. This requires domain knowledge, to specify what type of distributional changes are plausible. Formally, for a model $f(X)$ trained on data from $\mathbb{P}(X, Y)$, with loss function $\ell(f(X), Y)$, we seek

*Equal contribution, order determined by coin flip

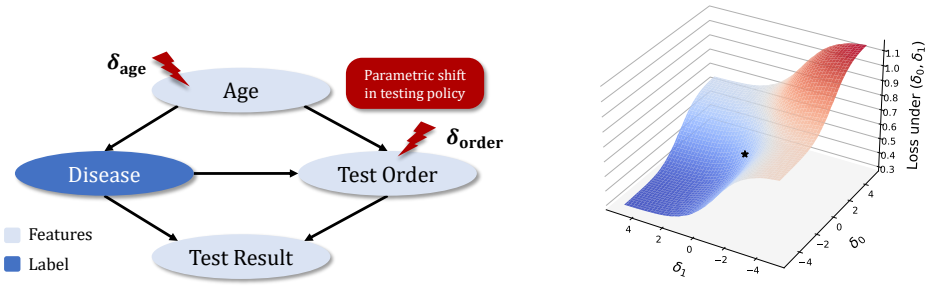


Figure 1.: (Left) Causal graph for Example 1. Our approach allows for simultaneous shifts in age and test ordering, parameterized by $\delta_{\text{age}}, \delta_{\text{order}}$. (Right) We illustrate a shift only in testing rates, using $s(Y; \delta_{\text{order}}) = \delta_1 \cdot Y + \delta_0(1-Y)$, where $\delta_{\text{order}} = (\delta_0, \delta_1)$. Here we plot the (non-concave) landscape of the expected cross-entropy loss of a fixed model over distributions parameterized by (δ_0, δ_1), with the training distribution given as the black star. Simulation details are given in Appendix B.1.

to understand the loss of the model under a set of *plausible* future distributions $\mathcal{P}$. We seek to evaluate the worst-case loss over $\mathcal{P}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(f(X), Y)], \tag{1}$$

and provide an interpretable description of a distribution P which maximizes this objective. If the value of the worst-case loss is low, this can build confidence prior to deployment, and otherwise, examining the worst-case distribution P can help identify weaknesses of the model. To illustrate, we use the following running example, inspired by Subbaswamy et al. [2021].

Example 1 (Changes in laboratory testing)**.** *We seek to classify disease* (Y) *based on the age* (A) *of a patient, whether a lab test has been ordered* (O)*, and test results* (L) *if ordered. The performance of a predictive model may be sensitive to changes in testing policies, as the fact that a test has been ordered itself is predictive of disease. Figure 1 (left) gives a plausible causal relationship between variables. Let* $\mathbb{P}(O|A,Y) = \sigma(\eta(A,Y))$*, where* σ *is the sigmoid function and* $\eta(A,Y)$ *is the log-odds. In Fig. 1 (right), we show the loss under a set of new distributions parameterized by* $\delta = (\delta_0, \delta_1)$*, where we modify* $\mathbb{P}_\delta(O|A,Y) = \sigma(\eta(A,Y) + s(Y;\delta))$ *for a shift function* $s(Y;\delta) = \delta_1 \cdot Y + \delta_0 \cdot (1-Y)$*, which modifies the log-odds of testing for both sick and healthy patients.*[1] *If* δ_0, δ_1 *are unconstrained, the worst-case occurs when all healthy patients are tested, and no sick patients are tested.*

The first challenge is to define a set of possible distributions $\mathcal{P}$ such that each distribution $P \in \mathcal{P}$ satisfies two desiderata: First, they should be *causally interpretable and simple to specify*, without placing unnecessary restrictions on the data-generating process. Second, they should be *realistic*, which often entails bounding the magnitude of the

[1] Code to reproduce figures and experiments in this paper can be found at https://github.com/clinicalml/parametric-robustness-evaluation

shift. We construct causally interpretable shifts by defining perturbed distributions $\mathbb{P}_\delta$ using changes in causal mechanisms, parameterized by a finite-dimensional parameter δ. Our main requirement is that the shifting mechanisms follow a conditional exponential family distribution. For discrete variables, this places no restriction on $\mathbb{P}$: In Example 1, O is binary and the log-odds $\eta(A, Y)$ can be any function of A, Y. We also demonstrate that constraining δ can ensure that shifts are realistic: The unconstrained worst-case shift in Example 1 is implausible, where all healthy patients (and no sick patients) are tested. (1) becomes

$$\sup_{\delta \in \Delta} \mathbb{E}_\delta[\ell(f(X), Y)], \tag{2}$$

where $\mathbb{E}_\delta$ is the expectation in the shifted distribution $\mathbb{P}_\delta$ and Δ is a bounded set of shifts.

The second challenge is evaluation of the expected loss under shift, as well as finding the worst-case shift. Under our definition of shifts, we show that the test distribution can always be seen as a reweighting of the training distribution, allowing for reweighting approaches, such as importance sampling, to estimate the expected loss under shifts. While this is practical for some distribution shifts, for others, importance sampling can lead to extreme variance in estimation. Further, finding the worst-case shift using a reweighted objective involves maximization over a non-concave objective (see Fig. 1), a problem that is generally NP-hard. We derive a second-order approximation to the expected loss under shift, and show how it can be estimated without the use of reweighting. For quadratic constraints Δ, we can approximate the general non-convex optimization problem in (2) with a non-convex, quadratically constrained quadratic program (QCQP) for which efficient solvers exist [Conn et al., 2000, Section 7]. We bound the approximation error of this surrogate objective, and show in experiments that it tends to find impactful adversarial shifts.

Our contributions are as follows:

1. We provide a novel formulation of robustness sets which are defined using parametric shifts. This formulation only require that the shifting mechanisms can be modelled as a conditional exponential family (see Section 2).

2. We derive a second-order approximation to the expected loss and provide a bound on the approximation error. We show that this translates the general non-convex problem into a particular non-convex quadratic program, for which efficient solvers exist (see Section 3).

3. In a computer vision task, we find that this approach finds more impactful shifts than a reweighting approach, while taking far less time to compute, and that the resulting estimates of accuracy are substantially more reliable (see Section 4).

Related work: A recent line of work learns predictive models that minimize objectives similar to (1), a task known as distributionally robust optimization [Duchi and Namkoong, 2021, Duchi et al., 2020, Sagawa et al., 2020]. Beyond our difference in motivation (we consider evaluation of a fixed model, not optimization), prior work typically

defines the robustness set using a notion of distance between distributions. In contrast, we consider sets of distributions arising from explicit parametric perturbations. Moreover, many of these approaches focus on changes in marginal distributions, while one of our primary motivations is handling conditional shifts. Closer to our work is Subbaswamy et al. [2021] who consider evaluating the loss under worst-case changes in a conditional distribution, but while we consider parametric shifts, they estimates the loss under worst-case $(1-\alpha)$ conditional subpopulation shifts. In some settings, such shifts may not represent plausible changes, as we demonstrate in Appendix B.4, where (in a simplified lab-testing example) the worst-case subpopulation is one where healthy patients are always tested, and sick patients never tested. Prior work on robustness to parametric interventions has been restricted to linear causal models with additive shift interventions [Rothenhäusler et al., 2021, Oberst et al., 2021, Kook et al., 2022]. Our work can be seen as extending those ideas to general non-linear causal models, where our focus is on evaluation rather than learning robust models. A separate line of work attempts to learn models with optimal worst-case performance under arbitrary causal interventions of various forms [Magliacane et al., 2018, Rojas-Carulla et al., 2018, Arjovsky et al., 2019, Subbaswamy et al., 2022] while we focus on explicitly bounded interventions that have potentially restricted parameterizations.

2. Defining Parametric Robustness Sets

Notation: Let $\mathbf{V}$ denote all observed variables, where $(X, Y) \subseteq \mathbf{V}$ for features X and labels Y, and let $\mathbb{P}(\mathbf{V})$ denote the training distribution. $\mathbb{E}[\cdot]$ and $\text{cov}(\cdot,\cdot)$ refer to the mean and covariance in $\mathbb{P}$, and for a shifted distribution $\mathbb{P}_\delta$ (Definition 1) we use $\mathbb{E}_\delta[\cdot]$, $\text{cov}_\delta(\cdot,\cdot)$. For a random variable Z, we use $\mathcal{Z}$ to denote the space of realizations, and d_Z for dimension e.g., $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$. For a set of random variables $\mathbf{V} = \{V_1, \ldots, V_d\}$, we use V_i to denote an individual element, and use $\text{PA}_\mathcal{G}(V_i)$ to denote the set of parents in a directed acyclic graph (DAG) $\mathcal{G}$, omitting the subscript when otherwise clear.

We begin with a general definition of a parameterized robustness set of distributions $\mathcal{P}$.

Definition 1. A *parameterized robustness set around* $\mathbb{P}(\mathbf{V})$ is a family of distributions $\mathcal{P}$ with elements $\mathbb{P}_\delta(\mathbf{V})$ indexed by $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$, with $0 \in \Delta$, where $\mathbb{P}_0(\mathbf{V}) = \mathbb{P}(\mathbf{V})$.

We give examples shortly that satisfy this general definition. To construct such a robustness set, we consider distributions $\mathbb{P}_\delta$ that differ from $\mathbb{P}$ in one or more conditional distributions (Assumption 1). We require that the relevant conditional distributions can be described by an exponential family.

Definition 2 (Conditional exponential family (CEF) distribution)**.** $\mathbb{P}(W|Z)$ is a conditional exponential family distribution if there exists a function $\eta(Z) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ such that the conditional probability density (for continuous W) or probability mass function (for discrete W) is given by

$$\mathbb{P}(W|Z) = g(W) \exp\Big(\eta(Z)^\top T(W) - h(\eta(Z))\Big), \tag{3}$$

where $T(W)$ is a vector of sufficient statistics, $T(W) \in \mathbb{R}^{d_T}$, $g(\cdot)$ specifies the density of a base measure and $h(\eta(Z))$ is a normalizing constant.

Definition 2 does not restrict $\mathbb{P}(W|Z)$ for binary/categorical W, and captures a wide range of distributions, including the conditional Gaussian (see Appendix B.2.1 for other examples). Definition 2 extends to marginal distributions where $Z = \varnothing$ and $\eta(Z)$ is a constant function.

Example 1 (Continued)**.** *Suppose the probability of ordering a test* (O) *depends on age* (A) *and disease* (Y)*, such that* $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$*, where* σ *is the sigmoid, and* η *is an arbitrary function. Here, Definition 2 is satisfied with* $W = O$*,* $Z = (A, Y)$*, and sufficient statistic* $T(O) = O$*.*

We now state our main assumption, where we distinguish between the terms in the joint distribution of $\mathbb{P}$ that shift, which we will need to model, and those that remain fixed, which we do not.

Assumption 1 (Factorization into CEF distributions)**.** *Let* $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ *be a "intervention set" of variables and let*

$$\mathbb{P}(\mathbf{V}) = \underbrace{\prod_{W_i \in \mathbf{W}} \mathbb{P}(W_i|Z_i)}_{\textit{Conditionals that shift}} \underbrace{\prod_{V_j \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_j|U_j)}_{\textit{Conditionals we do not model}} \tag{4}$$

be a factorization, where $Z_i, U_j, V_j \subseteq \mathbf{V}$ *are possibly overlapping sets of variables. We assume for each* W_i *that* Z_i *is known and that* $\mathbb{P}(W_i|Z_i)$ *satisfies Definition 2.*

If $\mathbb{P}(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$, the factorization in Assumption 1 is always satisfied by $Z_i = \mathrm{PA}_{\mathcal{G}}(W_i)$. Here, we require limited knowledge of the underlying graph, and only need to know the parents $\mathrm{PA}(W_i)$ for the variables W_i that shift. In Appendix B.2.2 we show that we can also consider shifts that extend Z_i to include additional variables, subject to an acyclicity constraint. We now define parametric perturbations and give the general form of the robustness sets that we consider in this work, involving simultaneous perturbations to multiple W_i.

Definition 3 (Parameterized shift functions and δ-perturbations)**.** Let $s(Z; \delta) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ be a *parameterized shift function* with parameters $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$ which is twice-differentiable with respect to δ and which satisfies $s(Z; 0) = 0$ for all Z. For $\mathbb{P}(W|Z)$ satisfying (3), we refer to

$$\mathbb{P}_\delta(W|Z) = g(W) \exp\Big(\eta_\delta(Z)^\top T(W) - h(\eta_\delta(Z))\Big)$$

as a δ-perturbation of $\mathbb{P}(W|Z)$ with shift function $s(Z; \delta)$, where $\eta_\delta(Z) := \eta(Z) + s(Z; \delta)$.

Example 1 (Continued)**.** *A model developer may be concerned about a uniform change in testing rates across all types of patients. This can be modelled by choosing* $s(Z; \delta) = \delta$*, for* $\delta \in \mathbb{R}$*, an additive intervention on the log-odds scale. A separate change in testing*

rates for sick and healthy patients could instead be modeled using $s(Z;\delta) = \delta_0(1-Y) + \delta_1 Y$*, using* $\delta \in \mathbb{R}^2$*. This reasoning extends readily to more complex shifts (e.g., allowing for age-specific changes in testing rates, with a non-linear dependence on age), as long as* $s(Z;\delta)$ *remains a parametric function.*

While the shift function $s(Z;\delta)$ is parametric, $\eta(Z)$ is unconstrained in Definitions 2 and 3. Note that this formulation includes multiplicative shifts $\eta_\delta(Z) = (1+\delta)\eta(Z)$ by letting $s(Z;\delta) = \delta \cdot \eta(Z)$.

Definition 4 (CEF parameterized robustness set)**.** For a distribution $\mathbb{P}$ and intervention set $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ satisfying Assumption 1, let each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ be a δ_i-perturbation (Definition 3) of $\mathbb{P}(W_i|Z_i)$. Then

$$\mathbb{P}_\delta(\mathbf{V}) = \Big(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\Big)\Big(\prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)\Big)$$

is called a δ-perturbation of $\mathbb{P}(\mathbf{V})$, and the robustness set $\mathcal{P}$ consists of all $\mathbb{P}_\delta$ for $\delta \in \Delta_1 \times \cdots \Delta_m$.

To estimate the expected loss under $\mathbb{P}_\delta$, we will typically[2] need to estimate $\eta(Z_i)$ for each $W_i \in \mathbf{W}$. However, we make no distributional assumptions on the remaining variables $\mathbf{V} \setminus \mathbf{W}$. This is useful in applications such as computer vision, where we do not need to restrict the generative model of images given attributes (e.g., background, camera type, etc), but can still model the expected loss under changes in the joint distribution of those attributes.

Remark 1 (Causal Interpretation of Shifts). If the DAG $\mathcal{G}$ represents a causal graph [Pearl, 2009], then $\mathbb{P}_\delta$ can interpreted as a change in causal mechanisms. We see this as an important perspective for interpreting and specifying plausible shifts, but our methods do not require a causal interpretation.

3. Evaluation of the Worst-Case Loss

For a fixed predictor and loss function, we can use data from $\mathbb{P}(\mathbf{V})$ to estimate the expected loss $\mathbb{E}_\delta[\ell] \coloneqq \mathbb{E}_\delta[\ell(f(X), Y)]$ for a fixed δ, and estimate the worst-case loss over all δ of bounded magnitude. In Section 3.1, we show that $\mathbb{P}_\delta$ shares support with $\mathbb{P}$, suggesting the use of reweighting estimators. However, these can exhibit high variance for shifts that produce large density ratios (see Appendix B.3.5 for an example), and moreover, maximizing a reweighted objective over δ is generally a non-convex problem. In Section 3.2 we derive an approximation to the expected loss under $\mathbb{P}_\delta$, yielding a tractable surrogate optimization problem under quadratic constraints such as $\|\delta\|_2 \leq \lambda$.

[2] As a special case, in Appendix B.3.2, we show the second-order approximation (Theorem 1) can be estimated in the case of variance-scaled mean-shifts in a conditional Gaussian without estimation of all of $\eta(Z)$.

Remark 2. The methods here can be used with an arbitrary predictor f and loss function $\ell := \ell(f(X), Y)$. We do not even require access to the original predictor f. Both methods here simply treat ℓ as a random variable in $\mathbb{P}$, for which we have samples from the training distribution.

3.1. Modelling shifted losses using reweighting

The shifts defined in Section 2 share common support, with the following density ratio.

Proposition 1. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 4,* $\operatorname{supp}(\mathbb{P}) = \operatorname{supp}(\mathbb{P}_\delta)$ *and the density ratio $w_\delta := \mathbb{P}_\delta / \mathbb{P}$ is given by*

$$w_\delta(\mathbf{V}) = \exp\left(\sum_{i=1}^{m} s_i(Z_i;\delta_i)^\top T_i(W_i)\right) \exp\left(\sum_{i=1}^{m} h(\eta_i(Z_i)) - h(\eta(Z_i) + s_i(Z_i;\delta_i))\right).$$

The proof can be found in Appendix B.6, along with all proofs for all other claims.

Example 1 (Continued)**.** *Suppose we perturb the probability of ordering a test O given age A and disease Y with shift function $s(Y;\delta) = \delta_0(1-Y)+\delta_1 Y$, independently changing the conditional probability of testing for healthy and sick patients. Here, the density ratio is given by*

$$w_\delta(O, A, Y) = \exp(s(Y;\delta)\cdot O)\frac{1+\exp(\eta(A,Y))}{1+\exp(\eta(A,Y)+s(Y;\delta))}. \tag{5}$$

To model the loss $\mathbb{E}_\delta[\ell]$ using data from $\mathbb{P}$, we can consider an importance sampling (IS) estimator [Horvitz and Thompson, 1952, Shimodaira, 2000], observing that $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(\mathbf{V})\cdot\ell]$. This requires estimation of the density ratio $w_\delta(\mathbf{V})$, and (given a sample $\{\mathbf{V}^j\}_{j=1}^n$ from $\mathbb{P}$) yields the estimator

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\mathrm{IS}} := \frac{1}{n}\sum_{j=1}^{n} \hat{w}_\delta(\mathbf{V}^j)\ell(\mathbf{V}^j). \tag{6}$$

In practice, (6) can have high variance when density ratios are large, and maximizing this equation with respect to δ is a general non-convex optimization problem, which is generally NP-hard to solve.

3.2. Approximating the shifted loss for exponential family models

We now propose an alternative approach for approximating the loss $\mathbb{E}_\delta[\ell]$. Recalling that $\mathbb{P}_{\delta=0} = \mathbb{P}$, we use a second-order Taylor expansion around the training distribution

$$\mathbb{E}_\delta[\ell] \approx \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\,\delta, \tag{7}$$

where $\mathbb{E}[\ell]$ denotes the loss in the training distribution and $\mathrm{SG}^1, \mathrm{SG}^2$ are defined as follows.

Definition 5 (Shift gradient and Hessian)**.** For a parametric shift satisfying Definition 1 where $\delta \mapsto \mathbb{E}_\delta[\ell]$ is twice-differentiable, we denote the *shift gradient* SG^1 and *shift Hessian* SG^2 as $\mathrm{SG}^1 := \nabla_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}$ and $\mathrm{SG}^2 := \nabla^2_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}$.

(7) is a local approximation of the loss, whose approximation error we bound in Theorem 2, with smaller approximation error for smaller shifts.[3] For $\mathbb{P}_\delta$ satisfying Definition 4, SG^1 and SG^2 can be computed as expectations in the training distribution, without estimation of density ratios. Recall that the conditional covariance is given by $\mathrm{cov}(A, B|C) := \mathbb{E}[(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])|C]$.

Theorem 1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 4, with intervened variables $\mathbf{W} = \{W_1, \ldots, W_m\}$ and shift functions $s_i(Z_i; \delta_i)$, where $\delta = (\delta_1, \ldots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top \,\mathrm{cov}\left(\ell,\, T_i(W_i)\middle| Z_i\right)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top \,\mathrm{cov}\left(\ell,\, \varepsilon_{T_i|Z_i}\varepsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \varepsilon_{T|Z}\right] & i = j \\ \mathrm{cov}(\ell,\; D_{i,1}^\top \varepsilon_{T_i|Z_i}\varepsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i; \delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\varepsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Theorem 1 handles arbitrary parametric shift functions in multiple variables, but for simple shift functions in a single variable, the notation simplifies substantially, as we show in Corollary 1.

Corollary 1 (Simple shift in a single variable)**.** *Assume the setup of Theorem 1, restricted to a shift in a single variable W, and that $s(Z; \delta) = \delta$. Then $D_1 = 1$, $D_2 = 0$, and*

$$\mathrm{SG}^1 = \mathbb{E}\left[\mathrm{cov}\left(\ell, T(W)\middle| Z\right)\right] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}\left[\mathrm{cov}\left(\ell, \varepsilon_{T|Z}\varepsilon_{T|Z}^\top\middle| Z\right)\right],$$

where $T(W)$ is the sufficient statistic of W and $\varepsilon_{T|Z} := T(W) - \mathbb{E}[T(W)|Z]$.

Example 1 (Continued)**.** *Suppose that age (A), which has no causal parents, follows a normal distribution with mean μ and variance σ^2, and that we wish to consider a shift in the mean. We can parameterize $\mathbb{P}(A)$ as an exponential family with parameter $\eta = \mu/\sigma$ and sufficient statistic $T(A) = A/\sigma$. Here, $s(\delta) = \delta$ implies a shift in the mean of δ*

[3] In Appendix B.3.3, we give an example of a linear-Gaussian generative model where this second-order expansion is exact, corresponding to the setting of anchor regression [Rothenhäusler et al., 2021].

standard deviations $\eta_\delta = \eta + s(\delta) = (\mu + \sigma\delta)/\sigma$*, and we can write that* $\mathrm{SG}^1 = \mathrm{cov}\,(\ell, A)\,/\sigma$ *and* $\mathrm{SG}^2 = \mathrm{cov}\left(\ell, (A - \mathbb{E}[A])^2\right)/\sigma^2$.

To estimate the shift gradient and Hessian from a sample from $\mathbb{P}$, for each $i = 1, \ldots, m$ we fit models $\hat{\mu}_\ell(Z_i) \approx \mathbb{E}[\ell|Z_i]$ and $\hat{\mu}_{W_i}(Z_i) \approx \mathbb{E}[T_i(W_i)|Z_i]$ and compute residuals on these predictions, which permits estimation of the gradient/Hessian as a sample average of residuals. A detailed treatment is given in Appendix B.3.1. Using these, we can estimate the expected loss as

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\mathrm{Taylor}} := \hat{\mathbb{E}}[\ell] + \delta^\top \hat{\mathrm{SG}}^1 + \frac{1}{2}\delta^\top \hat{\mathrm{SG}}^2 \delta. \tag{8}$$

Here, there are two sources of error: Finite-sample error, due to the estimates of $\mathrm{SG}^1, \mathrm{SG}^2$, as well as approximation error. The latter is bounded by the norm of δ and a term that depends on the covariance between the loss and the deviations of the sufficient statistic from its shifted mean.

Theorem 2. *Assume that* $\mathbb{P}_\delta, \mathbb{P}$ *satisfy the conditions of Theorem 1, with a shift in a single variable* W*, where* $s(Z;\delta) = \delta$*. Let* $E_{\delta,\,Taylor}$ *be the population Taylor estimate ((7)) and let* $\sigma(M)$ *denote the largest absolute value of the eigenvalues of a matrix* M*. Then*

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\,Taylor}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \sigma\left(\mathrm{cov}_{t\cdot\delta}(\ell, \varepsilon_{t\cdot\delta,T|Z}\varepsilon_{t\cdot\delta,T|Z}^\top) - \mathrm{cov}(\ell, \varepsilon_{0,T|Z}\varepsilon_{0,T|Z}^\top)\right) \cdot \|\delta\|^2,$$

where $T(W)$ *is the sufficient statistic of* $W|Z$ *and* $\varepsilon_{t\cdot\delta,T|Z} = T(W|Z) - \mathbb{E}_{t\cdot\delta}[T(W|Z)]$.

In exchange for considering a second-order approximation of the loss, we gain two benefits: Variance reduction and tractable optimization. First, as $\mathrm{SG}^1, \mathrm{SG}^2$ are not functions of δ, the variance of $\hat{E}_{\delta,\mathrm{Taylor}}$ is $O(\|\delta\|^4)$, while the variance of $\hat{E}_{\delta,\mathrm{IS}}$ can be much larger: We give a simple case in Appendix B.3.6 where $\mathrm{var}(\hat{E}_{\delta,\mathrm{Taylor}}) = O(\delta^4)$ while $\mathrm{var}(\hat{E}_{\delta,\mathrm{IS}}) = O(\delta^2 \exp(\delta^2))$. Second, maximizing $\hat{E}_{\delta,\mathrm{Taylor}}$ over the set $\|\delta\| \leq \lambda$ can be solved in polynomial time by exploiting the quadratic structure, while maximizing $\hat{E}_{\delta,\mathrm{IS}}$ over the constraints is generally hard, and may be infeasible in high dimensions.

3.3. Identifying worst-case parametric shifts

For $\lambda > 0$, we can locally approximate the worst-case loss over all distributions $\mathbb{P}_\delta$ where $\|\delta\|_2 \leq \lambda$ by finding the worst-case loss in the Taylor approximation

$$\sup_{\|\delta\|_2 \leq \lambda} \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta. \tag{9}$$

Since SG^2 is generally not negative definite, the maximization objective is non-concave. However, this particular problem is an instance of the 'trust region problem'[4] which is well-studied in the optimization literature [Conn et al., 2000], and can be solved in

[4]Not to be confused with the 'trust region *method*', which repeatedly solves the trust region *problem*.

polynomial time by specialized algorithms (see Pólik and Terlaky [2007, Section 8.1] for an example). This follows from the fact that strong duality holds, so that the optimal solution δ^* can be characterized in terms of the Karush-Kuhn-Tucker conditions [Boyd and Vandenberghe, 2004, Section 5.2]. For this problem, we use the `trsapp` routine from NEWUOA [Powell, 2006], as implemented in the python package `trustregion`. Depending on the application and prior knowledge, one may choose constraint sets that differ from $\|\delta\| \leq \lambda$. The strong duality of (9) holds for any quadratic constraint $\delta^\top A\delta + \delta^\top b \leq \lambda$, making it a general non-convex quadratically constrained quadratic program (QCQP), allowing for e.g., larger shifts in some directions than in others.

4. Experiments

4.1. Illustrative example: Laboratory testing

To build intuition, we illustrate our method in a simple case. Motivated by Example 1, we use a simple generative model where lab tests are more likely to be ordered (O) for sick patients (Y), and lab values (L) are predictive of Y.

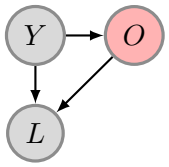


Figure 2.

$$Y \sim \mathsf{Ber}(0.5) \quad O|Y \sim \mathsf{Ber}(\sigma(\alpha + \beta Y)) \quad L|(Y, O = 1) \sim \mathcal{N}(\mu_y, 1)$$

where $\mu_1 = 0.5, \mu_0 = -0.5$, and we initialize with $\alpha = -1$, $\beta = 2$, so that $\mathbb{P}(O = 1|Y = 0) \approx 0.27$ and $\mathbb{P}(O = 1|Y = 1) \approx 0.73$, and the marginal probability of test ordering is $\mathbb{P}(O = 1) = 0.5$. When $O = 0$, we set L to a dummy value of $L = 0$. The underlying causal graph is given in Fig. 2. The predictive model $f(O, L)$ for this example is trained on data from $\mathbb{P}$ to predict Y using all available features: If lab tests are not available ($O = 0$), this model predicts Y based on the observed likelihood of Y given $O = 0$, and otherwise predicts using a logistic regression model trained on the cases where $O = 1$ in the training data.

Defining a shift function: $\mathbb{P}(O|Y)$ is a conditional exponential family with $\eta(Y) = \alpha + \beta Y$. We consider the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 models an overall change in testing rate, and δ_1 models an additional change in the likelihood of testing sick ($Y = 1$) patients.

Estimating the impact of shift using quadratic approximation: To start, we keep $\delta_1 = 0$ fixed and vary only δ_0, which uniformly increases or decreases testing. In Fig. 3, we show the ground-truth cross-entropy loss of $f(O, L)$ under perturbed distributions $\mathbb{P}_{\delta_0}$. We observe that the **direction** of the shift matters: In Fig. 3, the model performance slightly increases under a small increase in testing rates, but degrades if testing increases too much; moreover, the loss under shift is generally asymmetric, as a decrease hurts more than an increase in testing. In Fig. 3 (left), we demonstrate the use of the quadratic approximation described in Section 3.2. For illustration, we consider a robustness set of $\delta_0 \in [-2, 2]$, and see that the predicted worst-case shift coincides with the actual worst-case shift, and that the quadratic approximation is accurate for smaller values of δ.

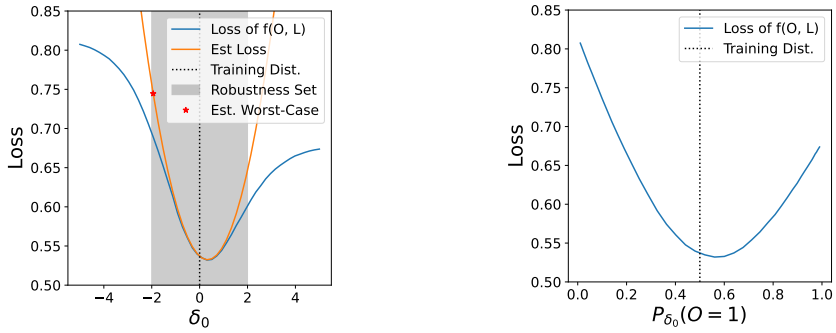


Figure 3.: The blue line gives the (unobserved) cross-entropy loss under parametric shifts, plotted with respect to the parameter δ_0 (left) and the resulting change in the marginal laboratory testing rate (right). We also provide the quadratic approximation (orange line), estimated using validation data, and the predicted worst-case shift (red star) for $|\delta_0| < 2$ (region in grey).

In Appendix B.4, we consider the case where δ_0 and δ_1 can both vary, and compare our approach to that of worst-case $(1 - \alpha)$ subpopulation shifts [Subbaswamy et al., 2021]. In the context of this example, we demonstrate that for any $1 - \alpha < 0.27$, the worst-case conditional subpopulation loss is achieved by having all healthy patients get tested, and no sick patients get tested. We contrast this with an iterative approach to designing constraints that is made possible by considering parametric shifts, where end-users can restrict the degree to which the shift differs across sick and healthy populations.

4.2. Detecting sensitivity to non-causal correlations

A predictive model may pick up on various problematic dependencies in the data that may not remain stable under dataset shift. To understand the impact of these dependencies, a model user may wish to understand which changes in distribution pose the greatest threats to model performance, and to measure the impact of these changes. To illustrate this use-case, we make use of the CelebA dataset [Liu et al., 2015], which contains images of faces and binary attributes (e.g., glasses, beard, etc.) encoding several non-causal features whose correlations may be unstable (e.g., the relation between gender and being bald). We consider the task of predicting gender (Y) from images of faces (X), and assess sensitivity to a shift in the distributions of attributes ($\mathbf{W}$).

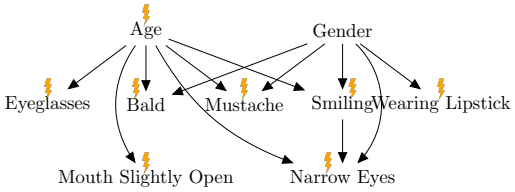


Figure 4.: Causal graph over attributes, where lightning bolts indicate changes in mechanisms.

Setup: To obtain ground-truth shifts in distribution, we generate synthetic datasets of faces using CausalGAN [Kocaoglu et al., 2018], trained on the CelebA data. We simulate attributes following the causal graph in Figure 4, and then simulate images

from the GAN conditioned on those attributes. We draw a training sample from this distribution $\mathbb{P}$, and fit a gender classifier $f(X)$ using the image data alone, by finetuning a pretrained ResNet50 classifier [Hu et al., 2018]. Each attribute W_i is binary, so we consider shifts in the log-odds $\eta_i(Z_i)$ of each attribute W_i given parents Z_i. Here, we use a maximally flexible shift function $s_i(Z_i; \delta_i) = \sum_{z \in \mathcal{Z}_i} \delta_{i,z} \mathbf{1}\{Z_i = z\}$, such that for $Z_i \in \{0,1\}^k$ there are 2^k parameters. Across all intervened variables, $\delta \in \mathbb{R}^{31}$. Due to the synthetic nature of our setup, we can simulate from $\mathbb{P}_\delta(X, \mathbf{W}, Y)$ to evaluate the ground-truth impact of this shift, simulating first from the shifted attribute distribution, and then simulating images from the GAN conditional on those attributes. We use the 0/1 loss $\ell = \mathbf{1}\{f(X) \neq Y\}$ and constrain δ by $\|\delta\|_2 \leq \lambda$ where $\lambda = 2$. We present results in terms of the accuracy, rather than the 0/1 loss itself.

Comparing importance sampling and Taylor across multiple simulations: We simulate $K = 100$ validation sets of size $n = 1{,}000$ from $\mathbb{P}$, in each estimating the worst-case shifts δ_{Taylor} (via the approach in Section 3.3) and δ_{IS}, where the latter corresponds to maximizing $\hat{E}_{\delta,\text{IS}}$ using a standard non-convex solver from the `scipy` library [Virtanen et al., 2020]. We simulate ground truth data ($n = 5{,}000$) from $\mathbb{P}_{\delta_{\text{IS}}}$ and $\mathbb{P}_{\delta_{\text{Taylor}}}$, to compare the two shifts. First, we demonstrate that the Taylor approach finds more impactful shifts. In Table 1 (right), we compare the average drop in accuracy using the Taylor shifts (between the validation and shifted test sets) and the IS shifts. For the former, the average drop is 3.8%, while for the latter, the average drop is 2.2%, a difference of 1.6%. In Fig. 5 (right) we plot the differences in test accuracy $\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}] - \mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$, showing that in 96% of cases, the Taylor method finds a more impactful shift. Second, when **only** used to evaluate the shift δ_{Taylor}, IS yields estimates $\hat{E}_{\delta_{\text{Taylor}},\text{IS}}$ of the test accuracy that are comparable to the Taylor-based estimates $\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$. However, while the optimal value of the Taylor objective tends to a reasonably accurate estimate of the shifted accuracy, the optimal value of the IS objective is a poor predictor. In Table 1 (right) we observe that the latter ($\hat{E}_{\delta_{\text{IS}},\text{IS}}$) is strongly biased in predicting $\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$. This bias leads to a large mean absolute prediction error (MAPE) of 0.069 (not shown in the table). This can be contrasted with a MAPE of 0.015 when using $\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$ to predict $\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}]$. Finally, we observe that the Taylor approach runs far faster, with an average run-time of $0.01s$ versus $2.14s$ for the IS approach.

Examining a single shift: To illustrate the type of shift found by our approach, we consider the δ_{Taylor} (of all those chosen over K validation sets) which yields the median test accuracy on the corresponding simulation from $\mathbb{P}_\delta$. We display the largest components of that δ in Table 1 (left). Among others, this shift entails a 5% increase in the probability of an older woman being bald, and a 5% decrease in the probability of a young woman wearing lipstick. This suggests that the learned classifier f relies on these associations in the images for prediction. We validate that this shift leads to a decrease in accuracy of around 3.8%, using simulated data from $\mathbb{P}_\delta$. To validate that this drop in accuracy is a non-trivial occurrence, we simulate $K = 400$ random shifts δ_k where $\|\delta_k\| = \lambda$ and evaluate the model accuracy in $\mathbb{P}_{\delta_k}$ using samples of size $n = 500$ (Fig. 5, left). As expected, our chosen shift δ yields a lower accuracy (red line) than all of the

Table 1.: (Left) Largest components of example shift δ where $\mathbb{P}$ and $\mathbb{P}_\delta$ denote conditional probabilities. (Right) Taylor and IS estimates vs. true accuracy at the shift δ found by Taylor approach, and IS estimate vs. true accuracy at shift δ found by IS approach.

Conditional		δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Bald	— Female, Old	0.899	0.047	0.109
Bald	— Male, Young	-0.800	0.378	0.214
Bald	— Male, Old	-0.680	0.622	0.455
Wearing Lipstick	— Female, Young	-0.618	0.924	0.868
Wearing Lipstick	— Female, Old	-0.543	0.953	0.921

	Example	Avg.
Acc. pre-shift ($\mathbb{E}[\mathbf{1}\{f(X) = Y\}]$)		0.912
Acc. post-shift ($\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}]$)	0.874	0.874
IS est. ($\hat{E}_{\delta_{\text{Taylor}},\text{IS}}$)	0.829	0.863
Taylor est. ($\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$)	0.844	0.863
Acc. post-shift ($\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$)	0.882	0.889
IS est. ($\hat{E}_{\delta_{\text{IS}},\text{IS}}$)	0.796	0.821

Figure 5.: (Left) Model accuracy at randomly drawn shifts. (Right) Difference in accuracy in the worst-case shifts identified by Taylor and importance sampling approaches. The Taylor method identifies a more adversarial shift than importance sampling in 96% of simulations (green).

random shifts.

5. Discussion

We argue for considering parametric shifts in distribution, to evaluate model performance under a set of changes that are interpretable and controllable. For parametric shifts in conditional exponential family distributions, we derive a local second-order approximation to the loss under shift. This approximation enables the use of efficient optimization algorithms (to find the worst-case shift), and empirically provides realistic estimates of the resulting loss. In a computer vision task, this approach finds more impactful shifts (in far less time) than optimizing a reweighted objective, and the estimates of shifted accuracy under the chosen shift are substantially more reliable.

That said, our approach is not without limitations. Our definition of parametric shifts and resulting approximation relies on the relevant mechanisms $\mathbb{P}(W|Z)$ being a conditional exponential family, and that the relevant variables are observed. As illustrated in our experiments, this can be used to model changes in the causal relationships **between** attributes of an image, but does not immediately extend to modelling changes in the distribution of images given a fixed set of attributes. As with any method that provides worst-case evaluation, there is potential for misuse and false confidence: If the specified shifts fail to capture important real-world changes, the resulting worst-case loss may

be overly optimistic and misleading. Even if used correctly, our approach examines a narrow measure of model performance, and a small worst-case error should not be used to claim that a model is free of problematic behavior. For example, implicit dependence on certain attributes (e.g., race in medical imaging [Banerjee et al., 2021]) may be problematic based on ethical grounds, even if it does not lead to major issues with predictive performance under small shifts in distribution.

Acknowledgments

We thank Jonas Peters, Chandler Squires, and Stefan Hegselmann for helpful feedback and discussion, and Irene Chen and Christina X Ji for providing comments on an earlier draft. MO and DS were supported in part by Office of Naval Research Award No. N00014-21- 1-2807. NT was supported by a research grant (18968) from VILLUM FONDEN.

Regularizing towards Causal Invariance: Linear Models with Proxies

Michael Oberst, Nikolaj Thams, Jonas Peters and David Sontag

Abstract

We propose a method for learning linear models whose predictive performance is robust to causal interventions on unobserved variables, when noisy proxies of those variables are available. Our approach takes the form of a regularization term that trades off between in-distribution performance and robustness to interventions. Under the assumption of a linear structural causal model, we show that a single proxy can be used to create estimators that are prediction optimal under interventions of bounded strength. This strength depends on the magnitude of the measurement noise in the proxy, which is, in general, not identifiable. In the case of two proxy variables, we propose a modified estimator that is prediction optimal under interventions up to a known strength. We further show how to extend these estimators to scenarios where additional information about the "test time" intervention is available during training. We evaluate our theoretical findings in synthetic experiments and using real data of hourly pollution levels across several cities in China.

1. Introduction

Ideally, predictive models would generalize beyond the distribution on which they are trained, e.g., across geographic regions, across time, or across individual users. However, models often learn to rely on signals in the training distribution that are not stable across domains, causing a drop-off in predictive performance. This problem is broadly known as dataset shift [Quionero-Candela et al., 2009].

Tackling this problem requires a formalization of how dataset shift arises, and how that shift impacts the conditional distribution of our target Y given features X. One way to formalize this shift is in terms of an underlying causal graph [Pearl, 2009], where changes between distributions are seen as arising from causal interventions on variables.

Conceptual example: In the causal graph given in Fig. 1, the variable A serves as a confounder. In a medical setting, A could represent smoking habits or socioeconomic status, which have a causal effect on current health status (X) as well as longer-term

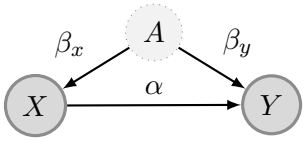


Figure 1.: Conceptual Example: A represents an (unobserved) socioeconomic variable, X represents current health status, and Y represents a long-term health outcome. All relationships are assumed to be linear, and coefficients are given. We consider a broader class of graphs in this work, see Fig. 2.

outcomes (Y). Importantly, A may not be recorded in our training data, and the distribution of A could vary across geography and time.

In the context of this causal graph, interventions which change the distribution of A will also alter the conditional mean $\mathbb{E}(Y \mid X)$. Under the linear relationships in Fig. 1, the optimal least-squares predictor $\hat{Y} = \gamma^* X$ under the test distribution depends on the test-time variance in A, in that

$$\gamma^* = \begin{cases} \alpha, & \text{if after intervention } A = 0 \\ \alpha + \frac{\beta_Y}{\beta_X}, & \text{if after intervention } \operatorname{var}(A) \to \infty. \end{cases}$$

The first predictor encodes the direct causal effect of X on Y, but is only optimal in the setting where the correlations induced by A are removed by fixing it to a constant value of zero (the same holds when including intercepts and allowing for non-zero means). The second predictor, on the other hand, renders the distribution of the residual $Y - \hat{Y}$ independent of A, and is therefore robust to arbitrary interventions upon A. However, this is only optimal under arbitrarily strong interventions on A.

Balancing performance and invariance: Instead of seeking an invariant predictor that is robust to arbitrary interventions on A (like the second predictor above), we instead seek to minimize a worst-case loss under bounded interventions of a given strength. We contrast this with work that seeks to discover causal relationships as a route to invariance [Rojas-Carulla et al., 2018, Magliacane et al., 2018], optimize for invariance directly across environments [Arjovsky et al., 2019], or use known causal structure to select predictors with invariant performance [Subbaswamy et al., 2019].

Our proposed objective takes the form of a standard loss, plus a regularization term that encourages invariance. This builds upon Rothenhäusler et al. [2021], who introduce a similar objective, and prove that their objective optimizes a worst-case loss over bounded interventions on A, under a large class of linear structural causal models.

In contrast to Rothenhäusler et al. [2021], we do not assume that A is observed. Instead we assume that, during training, we have access to noisy proxies of A. For most of the paper, we assume that neither A nor proxies are available during testing. With this in mind, our contributions are as follows

- *Distributional robustness to bounded shifts*: In Section 3, we show that a single proxy can be used to construct estimators with distributional robustness guarantees under bounded interventions on A. However, these estimators are robust to

a strictly smaller set of interventions, compared to when A is used directly, and the size of this set depends on the (unidentifiable) noise in the proxy. When two proxies are available, we propose a modified estimator that can be used to recover the same guarantees as when A is observed.

- *Targeted shifts*: In Section 4, we show how to target our loss to interventions on A contained in a specified robustness set. We show that this formulation includes Anchor Regression as a special case, but also allows for sets that are not centered around the mean of A. In this setting we give an estimator, using two proxies, that identifies the target loss.

In Section 5, we evaluate our theoretical findings on synthetic experiments, and in Section 6 we demonstrate our method on a real-world dataset consisting of hourly pollution readings across five major cities in China.

2. Preliminaries

2.1. Notation

We use upper case letters X to denote (possibly vector-valued) random variables, and lower-case letters x to denote values in the range of those random variables. Vectors are assumed to be column vectors, so that $X \in \mathbb{R}^{d_X}$ indicates that $X = (X_1, \ldots, X_{d_X})^\top$, a column vector of d_X random variables. We use $\Sigma_X \in \mathbb{R}^{d_X \times d_X}$ to denote the covariance matrix of a variable X. We use bold upper-case letters $\mathbf{X}$ to denote a data matrix in $\mathbb{R}^{n \times d_X}$, consisting of n i.i.d. observations of X, and $\mathbf{1}\{\cdot\}$ as an indicator random variable. When dealing with matrices C, D, we use $C \prec D$ and $C \preceq D$ to indicate the positive definite and positive semi-definite partial order, respectively. That is, $C \prec D$ if $D - C$ is positive definite (PD), and $C \preceq D$ if $D - C$ is positive semi-definite (PSD). We use Id to denote the identity matrix, whose dimension is given by context. All proofs are provided in the supplementary material.

2.2. Linear structural causal model

We assume the general class of causal graphs represented in Fig. 2, where $X \in \mathbb{R}^{d_X}$ denotes observed covariates that can be used in prediction, $Y \in \mathbb{R}^{d_Y}$ is the target we seek to predict, $H \in \mathbb{R}^{d_H}$ are unobserved variables, and $A \in \mathbb{R}^{d_A}$ represents anchor variables, which are assumed to have no causal parents in the graph. We assume the linear structural causal model (SCM) given in Assumption 1.

Assumption 1 (Linear SCM)**.** *We assume the SCM*

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \varepsilon, \tag{1}$$

where A, ε have zero mean, bounded covariance, and are independently distributed. We assume that $\mathbb{E}[AA^\top]$ and $\mathrm{Id} - B$ are invertible, where Id *is the identity matrix. See Fig. 2*

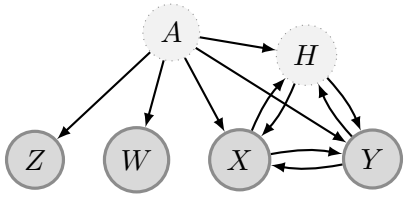


Figure 2.: In contrast to Rothenhäusler et al. [2021], we assume that anchor variables (denoted A) are unobserved, but that we have access to either one or two proxies W, Z. Observed variables are shown in dark grey and unobserved variables in light grey. We do not assume knowledge of the causal structure between A, X, H, Y (except that A has no causal parents). The relationship between X, H, Y could be cyclic, but all relationships are linear.

for a graphical representation.

Note that we do not assume here (or anywhere in this paper) that either A or ε is Gaussian. The invertibility of $\mathrm{Id} - B$ is satisfied if the causal graph is a directed acyclic graph. The matrices B, M_A encode the linear causal relationships. For instance, Fig. 1 can be represented in this form by $B = \begin{bmatrix} 0 & 0 \\ \alpha & 0 \end{bmatrix}$, $M = \begin{bmatrix} \beta_X \\ \beta_Y \end{bmatrix}$. In general, $\varepsilon \in \mathbb{R}^D$, $B \in \mathbb{R}^{D \times D}$, and $M \in \mathbb{R}^{D \times d_A}$, where $D := d_X + d_Y + d_H$. We assume that $d_Y = 1$ for simplicity.

2.3. Distributional robustness of anchor regression

Our goal is to learn a predictor $f^*(X)$ of Y that minimizes a worst-case risk of the following form

$$f^* = \arg\min_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Y, f(X))], \tag{2}$$

where $\mathcal{F}$ denotes a hypothesis class of possible predictors, $\mathcal{P}$ denotes a set of possible distributions, and ℓ represents our loss function. We take the class $\mathcal{P}$ to consist of distributions that arise as the result of causal interventions on A, and seek to learn a linear predictor to minimize mean-squared error.

We use $\mathbb{P}$ to refer to the observational distribution, and $\mathbb{P}_{do(A:=\nu)}$ to refer to the distribution under interventions on A, where the variable A is replaced by the random variable ν, and ν is assumed to be independent of the noise vector ε. We often write

$$R(\gamma) := Y - \gamma^\top X$$

as a random variable that represents the residual of a predictor $\gamma \in \mathbb{R}^{d_X}$. Importantly, Assumption 1 implies that for any γ, $\mathbb{E}[R(\gamma) \mid A]$ can be written as a linear function in A.

In this setting, Rothenhäusler et al. [2021] propose the following objective, defined here with respect to the observational distribution $\mathbb{P}$ (rather than a finite sample)

Definition 1 (Anchor Regression)**.**

$$\ell_{AR}(A;\gamma,\lambda) \coloneqq \ell_{LS}(X,Y;\gamma) + \lambda \ell_{PLS}(X,Y,A;\gamma), \tag{3}$$

where $\lambda \geq -1$ is a hyperparameter and

$$\ell_{LS}(X,Y;\gamma) \coloneqq \mathbb{E}\left[R(\gamma)^2\right] \tag{4}$$

$$\ell_{PLS}(X,Y,A;\gamma) \coloneqq \mathbb{E}\left[\left(\mathbb{E}\left[R(\gamma) \mid A\right]\right)^2\right]. \tag{5}$$

The first term ℓ_{LS} encodes the least-squares objective, while the second term ℓ_{PLS} encodes the residual error which can be predicted from A, which we refer to as the projected least-squares error. For $\lambda > 0$, the second term adds an additional penalty (beyond that of ordinary least squares) when the bias varies across values of A. The second term (5) can also be written in the linear setting of Assumption 1 as

$$\ell_{PLS}(A;\gamma) = \mathbb{E}[R(\gamma)A^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AR(\gamma)^\top], \tag{6}$$

where we drop the dependence on X, Y for notational simplicity. Under Assumption 1, (3) corresponds to a worst-case loss under distributional shift caused by bounded intervention on A [Rothenhäusler et al., 2021, Theorem 1]

$$\ell_{AR}(A;\gamma,\lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2], \tag{7}$$

where the robustness set is given by

$$C_A(\lambda) \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]\}. \tag{8}$$

Since minimizing ℓ_{AR} is equivalent to ordinary least squares (OLS) regression when $\lambda = 0$, this also provides a natural robustness guarantee for the OLS estimator, where $C_{OLS} \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\}$. In an identifiable instrumental variable setting, the minimizer converges against the causal parameter for $\lambda \to \infty$ [e.g. Jakobsen and Peters, 2021, eq. (71)]; the ℓ_{PLS} term has therefore been referred to as 'causal regularization' [e.g. Bühlmann and Ćevid, 2020], and has also been denoted by ℓ_{IV} [Rothenhäusler et al., 2021], as $\operatorname{cov}(A, R(\gamma)) = \mathbf{0}$ if and only if $\ell_{PLS}(\gamma) = 0$.

3. Distributional Robustness to Bounded Shifts

We first assume the existence of a noisy proxy W, conditionally independent of (X, Y, H) given A (see Fig. 2).

Assumption 2 (Single proxy with additive noise)**.** *In the context of Assumption 1, W is generated as follows*

$$W \coloneqq \beta_W^\top A + \varepsilon_W,$$

where ε_W has mean zero, bounded covariance, and is independent of (A, ε). In addition, we assume that the second moment matrix $\mathbb{E}[WW^\top]$ is invertible.

Under mild identifiability conditions (e.g., that β_W is full rank) one can show (see Appendix C.3.2) that

$$\ell_{PLS}(A;\gamma) = 0 \iff \ell_{PLS}(W;\gamma) = 0, \tag{9}$$

Hence, a single proxy is enough (in the population case) to identify whether the sharp constraint $\ell_{PLS}(\gamma) = 0$ holds, representing invariance to interventions of arbitrary strength. This corresponds to the fact that if A is a valid instrumental variable, then so is W [Hernán and Robins, 2006].

However, we consider interventions on A that are not of arbitrarily large strength. With that in mind, in Section 3.1, we demonstrate that (i) when a single proxy W is used in place of A, a robustness guarantee holds, but the robustness set is reduced relative to (8), (ii) the extent of this reduction depends on the signal-to-variance relationship in W, and (iii) this relationship is not generally identifiable from the observational distribution over (X, Y, W) alone. In Section 3.2, we show that in the setting where two proxies are available, the same guarantees as for an observed A can be obtained. We do so constructively, giving a regularization term whose population version is equal to $\ell_{PLS}(A;\gamma)$.

3.1. Robustness with a single proxy

First, we establish the robustness set of Anchor Regression when a single proxy is used in place of A. We refer to this as Proxy Anchor Regression, to distinguish it from the case when A is observed, but the only difference from Definition 1 is that W is used in place of A.

Definition 2 (Proxy Anchor Regression)**.** Let ℓ_{LS}, ℓ_{PLS} be defined as in (4) and (6). We define

$$\ell_{PAR}(W;\gamma,\lambda) \coloneqq \ell_{LS}(\gamma) + \lambda \ell_{PLS}(W;\gamma), \tag{10}$$

where $\lambda \geq -1$ is a hyperparameter and we suppress the dependence on X, Y in the notation.

Theorem 1. *Under Assumptions 1 and 2, for all $\gamma \in \mathbb{R}^{d_X}$ and for all $\lambda \geq -1$*

$$\ell_{PAR}(W;\gamma,\lambda) = \sup_{\nu \in C_W(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2],$$

where the robustness set is given by

$$C_W(\lambda) \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda \Omega_W\} \tag{11}$$

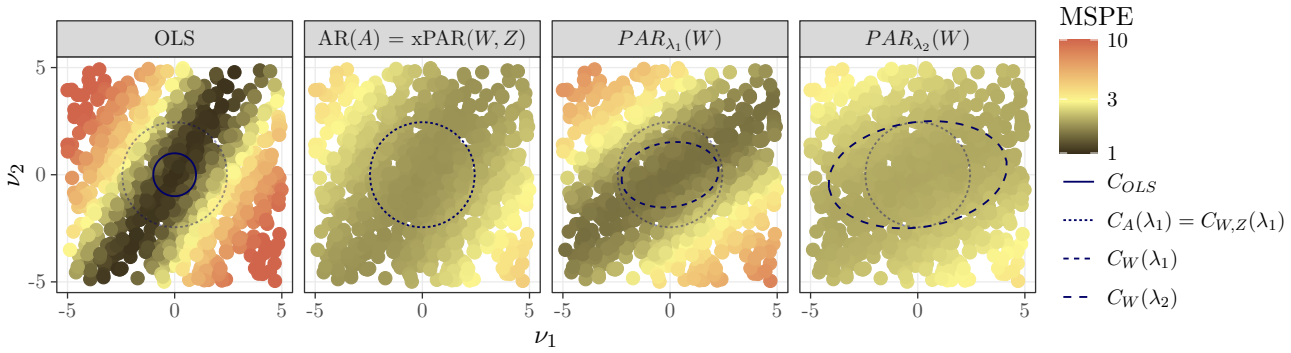


Figure 3.: Test performance under interventions $do(A := (\nu_1, \nu_2))$ which give rise to different test distributions over X and Y. Each dot corresponds to a different intervention (i.e., test distribution on X, Y), and the color gives the resulting mean squared prediction error (MSPE). **(Far Left)** OLS performs well for interventions in the set C_{OLS} (solid circle), corresponding to the training covariance of A. However, it performs poorly under interventions far from this region (e.g., top left). **(Middle Left)** Anchor Regression (AR) minimizes the worst-case loss over interventions on A within the region $C_A(\lambda_1)$ (cf., (8)), a rescaling of C_{OLS}. There is a trade-off, with better performance than OLS under large interventions, but worse performance under small interventions. Given two proxies W, Z, we introduce Cross-Proxy Anchor Regression (xPAR, cf., (14)) and prove that it minimizes the same worst-case loss. **(Middle Right)** When only a single proxy W is used in place of A, the result is a weaker guarantee, in the form of a smaller robustness set $C_W(\lambda_1)$ (cf., (11)) for the same value of λ_1. The shape of this set depends on the noise in the proxy along different dimensions. **(Far Right)** As a result, there does not generally exist a λ_2 such that $C_W(\lambda_2) = C_A(\lambda_1)$. If we choose some $\lambda_2 > \lambda_1$ such that $C_A(\lambda_1) \subset C_W(\lambda_2)$, we enforce a stronger constraint than intended, resulting in an unwanted trade-off between performance and robustness.

and where Ω_W *is defined as*

$$\Omega_W := \mathbb{E}[AW^\top]\Big(\mathbb{E}[WW^\top]\Big)^{-1}\mathbb{E}[WA^\top]. \tag{12}$$

Intuitively, Ω_W defines a signal-to-variance relationship in W, and this determines the robustness guarantee. In the case where both $A, W \in \mathbb{R}$ are one-dimensional, and A has unit variance, the robustness sets simplify to

$$\begin{aligned} C_{OLS} &= \{\nu : \mathbb{E}[\nu^2] \leq 1\} \\ C_W(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda \cdot \rho_W\} \\ C_A(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda\}, \end{aligned}$$

where $\rho_W := \beta_W^2/(\beta_W^2 + \mathbb{E}\varepsilon_W^2) < 1$ is the signal-to-variance ratio of W, also referred to as the reliability ratio in the measurement error literature [Fuller, 1987]. Thus, in the one-dimensional case, the robustness set using W is strictly smaller than the one obtained by using A when $\lambda > 0$, except in the case where $\varepsilon_W = 0$ a.s. This result generalizes to higher dimensions.

Proposition 1. *Assume Assumptions 1 and 2 and that* $\mathbb{E}[\varepsilon_W \varepsilon_W^\top] \in \mathbb{R}^{d_W \times d_W}$ *is positive definite. Then for* $\lambda > 0$

$$C_{OLS} \subseteq C_W(\lambda) \subset C_A(\lambda),$$

and the set $C_W(\lambda)$ *increases monotonically when* $\mathbb{E}[\varepsilon_W \varepsilon_W^\top]$ *decreases w.r.t. the partial matrix ordering. If* $d_W = d_A$, β_W *is full rank, and* $\varepsilon_W = 0$ *a.s., then* $C_W(\lambda) = C_A(\lambda)$.

If Ω_W were known, we could choose a larger λ^* such that $C_A(\lambda) \subseteq C_W(\lambda^*)$. In contrast to the one-dimensional case, where we could choose $\lambda^* = \lambda/\rho_W$ to obtain an equality $C_A(\lambda) = C_W(\lambda^*)$, we cannot generally achieve equality in higher dimensions (see Fig. 3).

However, Ω_W is not generally identifiable from the observed distribution over (X, Y, W) alone. Moreover, SCMs compatible with the observed distribution react differently under interventions on A and yield different coefficients that are optimal w.r.t. interventions in $C_A(\lambda)$. Consequently, in this setting, it is not possible to recover the guarantees of Anchor Regression without further assumptions (e.g., on Ω_W). See Appendix C.2 for an example.

Note that these results apply regardless of whether or not β_W is full rank. However, if β_W is not full rank, then there will be directions of variation in A that are not reflected in W, and we will not be able to achieve additional robustness (beyond that of OLS) against interventions along these directions.

3.2. Robustness with two proxies

We now show that if we have two (sufficiently different) proxies for A, then it is possible to recover the original robustness set using a different regularization term. We denote these proxies by W, Z, as shown in Fig. 2. In this setting, the structural causal model over (X, Y, H, A) can still be written in the form of Equation (1), where we make the following additional assumptions.

Assumption 3 (Proxies with additive noise)**.** *In the context of Assumption 1,* Z, W *are generated as follows*

$$W := \beta_W^\top A + \varepsilon_W \qquad \text{and} \qquad Z := \beta_Z^\top A + \varepsilon_Z,$$

where $\varepsilon_W, \varepsilon_Z$ *are mean-zero with bounded covariance, and* $\varepsilon_W, \varepsilon_Z, \varepsilon, A$ *are jointly independent.*

Assumption 4. *The dimensions of* A, W, Z *are equal,* $d_A = d_W = d_Z$, *and* β_W, β_Z *are full-rank.*

Note that Assumption 4 also implies that the second moment matrix $\mathbb{E}[ZW^\top]$ is invertible.

To build intuition, note that this assumption is trivially satisfied in the setting where $W = A + \varepsilon_W$ and $Z = A + \varepsilon_Z$, i.e., where W and Z are two noisy observations of A. More generally, Assumption 4 rules out directions of variation in A that are undetectable in W or Z.

In this setting we introduce the following loss, and prove that it is equal to the worst-case loss obtained when A is observed (c.f., (7))

Definition 3 (Cross-Proxy Anchor Regression)**.**

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) := \ell_{LS}(X, Y; \gamma) + \lambda \ell_\times(W, Z; \gamma),$$

where we refer to

$$\ell_\times(W, Z; \gamma) := \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZR(\gamma)^\top], \tag{13}$$

as the cross-proxy regularization term.

Theorem 2. *Under Assumptions 1, 3 and 4, for any $\gamma \in \mathbb{R}^{d_X}$ and any $\lambda \geq -1$*

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2], \tag{14}$$

where $C_A(\lambda) = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1 + \lambda)\mathbb{E}[AA^\top]\}$.

$\ell_{\times PAR}$ is convex in γ and has a closed form solution for its minimizer based only on the population moments of X, Y, W and Z (see Proposition C.4 in the supplement).

To build intuition for why Assumption 4 is required for this result, consider an example where W, Z are both scalars ($d_W = d_Z = 1$) and A has two independent dimensions (A_1, A_2). In this example, if both proxies measure the same dimension A_1, then variation in A_2 is not detectable in either proxy, and we cannot optimize for robustness to interventions on A_2. On the other hand, if W only measures A_1 (e.g., $W = A_1 + \varepsilon_W$), and Z only measures A_2 (e.g., $Z = A_2 + \varepsilon_Z$), then we cannot use Z to identify the signal-to-variance ratio of W, and vice-versa. In this case, (W, Z) is effectively a single two-dimensional proxy in the framework of Section 3.1, where we showed that recovering the guarantees of Anchor Regression is not generally possible. Intuitively, we need all directions of variation in A to have some influence on both proxies (i.e., β_W, β_Z full rank), and hence require that W, Z have sufficiently large dimension.

4. Targeted Anchor Regression: Incorporating Additional Shift Information

We now generalize Anchor Regression to an estimator that is targeted to be robust against particular shifts, and demonstrate that we can similarly handle this setting when only proxies of A are observed. In Section 2.3 we showed that Anchor Regression

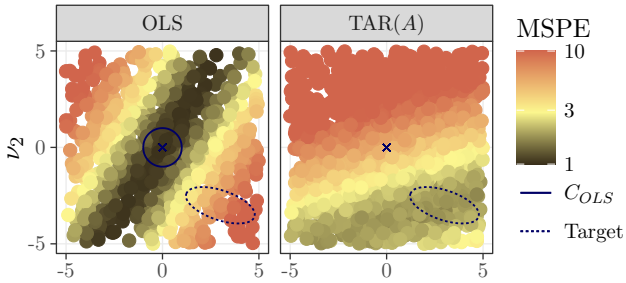


Figure 4.: Targeted Anchor Regression allows for minimizing the worst-case loss in regions (dashed ellipse) that may differ in location, size, and shape from the regions in Fig. 3 (OLS copied for reference). Every point ν represents a test distribution $do(A := \nu)$, the color indicating the mean squared prediction error in this distribution. Cross marks the origin. The TAR estimator achieves its minimal test loss at the center of the targeted region.

minimizes the worst-case loss over the set $C_A(\lambda)$ of all interventions $do(A := \nu)$ where $\mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]$. For deterministic ν, $C_A(\lambda)$ is an ellipsoid centered at 0, and its width in each direction is proportional to the variation of A in that direction. However, we may desire a different robustness set: For instance, if we anticipate a particular shift μ_ν in the mean of A, or if we want to add extra protection against particular directions of variation in A. This can be formalized as a robustness set defined by an ellipsoid that may not be centered at 0, nor be proportional to $\mathbb{E}[AA^\top]$. The estimator developed in this section can incorporate such prior beliefs.

More formally, instead of considering robustness against interventions $do(A := \nu)$ over the set $\nu \in C_A(\lambda)$, we now assume that we have additional information on the nature of ν, which is specified in the form of a vector μ_ν and a symmetric PSD matrix Σ_ν. We introduce a new method, Targeted Anchor Regression, minimizing what we refer to as the *targeted loss*. We prove in Propositions 2 and 3 that minimizing this objective can be interpreted in two ways: First, as minimizing an expected loss over interventions ν with a known mean and covariance, or minimizing a worst-case loss over deterministic interventions ν contained in an ellipsoid robustness set (as discussed above). This is visualized in Fig. 4.

4.1. Targeting when A is observed

We first consider the case when A is observed during training, and the mean and covariance of ν are known, given by μ_ν, Σ_ν. Importantly, for a given γ we have $\mathbb{E}[R(\gamma) \mid A = a] = b_\gamma^\top a$, where, writing $\Sigma_A := \mathbb{E}[AA^\top]$,

$$b_\gamma^\top := \mathbb{E}[R(\gamma)A^\top]\Sigma_A^{-1}. \tag{15}$$

Definition 4 (Targeted Anchor Regression)**.** Let $\mu_\nu \in \mathbb{R}^{d_A}$, and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$, where

Σ_ν is a symmetric PSD matrix.

$$\begin{aligned}&\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)\\ &:= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma + \left(b_\gamma^\top \mu_\nu - \alpha\right)^2,\end{aligned} \tag{16}$$

where b_γ is defined in (15), and Σ_A is the covariance of A.

Proposition 2. *Under Assumption 1, and the assumption that $\nu \perp\!\!\!\perp \varepsilon$, we have, for all $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,*

$$\ell_{TAR}(A; \mu_\nu, \Sigma_\nu; \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],$$

where $\mu_\nu = \mathbb{E}[\nu]$ and Σ_ν is the covariance matrix of ν.

Importantly, the objective in (16) is convex in (γ, α), and has a closed-form solution (see Proposition C.5 in the supplement). If ν is a known constant, then this corresponds to performing OLS using both X and A as predictors during training, and using the known value of ν for A for prediction (see Appendix C.3.3.2). However, if for example ν exhibits more variance than A along certain directions, and less variance along others, then the targeted regression parameter differs from standard solutions. Optimizing the objective in (16) can also be interpreted as optimizing a worst-case loss over interventions $do(A := \nu)$ in a certain set.

Proposition 3. *Under Assumption 1, we have, for all $\mu_\nu \in \mathbb{R}^{d_A}$ and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$ being a symmetric positive definite matrix, that*

$$\begin{aligned}&\arg\min_{\gamma,\alpha} \ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)\\ &\quad = \arg\min_{\gamma,\alpha} \sup_{\nu \in T(\mu_\mu, \Sigma_\nu)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],\end{aligned}$$

where the supremum is taken over (deterministic or random) shifts ν of the form $\nu = \mu_v + \delta$, where δ satisfies the constraint that $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$. If δ is random, we require that it is independent of all other random variables. In other words, we can write that ν lies in the set

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : \mathbb{E}[(\nu - \mu_\nu)(\nu - \mu_\nu)^\top] \preceq \Sigma_\nu\}.$$

Note that the expectation in the constraint T is with respect to the random variable ν. This covers the case in which ν (and hence δ) is deterministic, in which case it is equal to a fixed value with probability one.

Proposition 3 shows that Targeted Anchor Regression generalizes Anchor Regression to a broader class of robustness sets, that need not depend explicitly on $\mathbb{E}[AA^\top]$. In particular, Anchor Regression can be viewed as a special case, where $\Sigma_\nu = (1 + \lambda)\Sigma_A$ and $\mathbb{E}[\nu] = 0$, in which case the objectives are equal for $\alpha = 0$. In the following, we adopt the interpretation of μ_ν, Σ_ν as specifying a mean and covariance of ν (Proposition 2).

4.2. Targeting with proxies

In the single-proxy setting, we define Proxy Targeted Anchor Regression as using W in place of A in (16). We assume a known mean and covariance of W under $\mathbb{P}_{do(A:=\nu)}$, used in place of μ_ν, Σ_ν. By similar arguments to those in Section 3.1, this approach does not generally yield the optimal predictor, in a way that depends on the (unidentified) signal-to-variance relationship in W. Given the similarity, we defer details to Appendix C.4.

When two proxies W, Z are available, we can recover the statement from Proposition 2 using a modified estimator, by similar arguments to those in Section 3.2. The core observation is that we can construct a linear term

$$a_\gamma^\top := \mathbb{E}[R(\gamma)Z^\top](\mathbb{E}[WZ^\top])^{-1}, \tag{17}$$

which, if $\beta_Z = \beta_W = \mathrm{Id}$ can be seen as a linear IV estimate of $b_\gamma^\top$ in (15), an estimator used in the measurement error literature given repeated noisy measurements of a single variable [Fuller, 1987]. In our case, (17) identifies $b_\gamma^\top$ only up to the linear transformation β_W, but this is sufficient to identify the targeted loss.

Definition 5 (Cross-Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} \in \mathbb{R}^{d_W}$, and $\tilde{\Sigma}_W \in \mathbb{R}^{d_W \times d_W}$, where $\tilde{\Sigma}_W$ is a symmetric positive semi-definite matrix. We define

$$\begin{aligned}&\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha)\\ &:= \ell_{LS}(\gamma) + a_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W\right) a_\gamma + \left(a_\gamma^\top \tilde{\mu} - \alpha\right)^2,\end{aligned}$$

where a_γ is defined in (17).

In Theorem C.1 (Appendix C.4) we prove, analogous to Theorem 2, that this population objective is equal to that of Targeted Anchor Regression (16).

5. Synthetic Experiments

In Section 5.1, we show that Cross-Proxy Anchor Regression (xPAR) outperforms Proxy Anchor Regression (PAR) in settings with noisy proxies. As the noise increases, xPAR continues to match Anchor Regression (AR) test performance under intervention, while PAR approaches OLS. In Section 5.2, we demonstrate the risks of attempting to correct for this noise by assuming a certain signal-to-variance ratio. In Section 5.3 we demonstrate another benefit of xPAR over PAR, giving an example where it places more weight on causal predictors relative to PAR. Finally, in Section 5.4, we highlight the trade-off between using Targeted Anchor Regression (TAR) vs. OLS and AR, showing that TAR improves performance under the targeted shift, at the cost of incurring additional error on the training distribution. Code for experiments is available at `https://github.com/clinicalml/proxy-anchor-regression`.

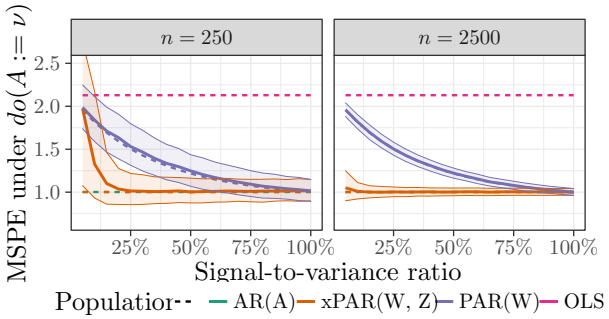


Figure 5.: Mean squared prediction error (MSPE) under interventions $do(A := \nu)$ for estimators PAR and xPAR. We display population losses for the population parameters as dashed lines, and median empirical MSPE when fit from data as solid lines, with shaded regions covering the 25% to 75% quantiles.

5.1. Mean squared prediction error under intervention

We demonstrate on synthetic data that xPAR recovers similar test performance to AR, while the performance of PAR degrades as the signal-to-variance ratio (SVR) of the proxies decreases. We simulate training data (at different levels of signal-to-variance) from an SCM with the structure given in Fig. 2, fix $\lambda := 5$ and fit PAR and xPAR. We then choose a fixed intervention ν, and simulate test data under the intervened distribution, evaluating our learned predictors.

In Fig. 5, we see that the test errors for xPAR and AR coincide (see Theorem 2) while PAR interpolates between OLS and AR, depending on the signal-to-variance ratio (see Proposition 1). Appendix C.5 gives additional implementation details on this and remaining experiments.

5.2. Misspecified signal-to-variance ratio

In Section 3.1, we noted that if the (unidentified) signal-to-variance ratio (SVR) were known, we could correct for it when using PAR with a single proxy. Here we demonstrate the implications of incorrectly specifying this correction. We simulate data from the same SCM as in Section 5.1, with varying (true) signal-to-variance ratio.

In Fig. 6, for the predictor chosen by PAR, we plot the estimated worst-case MSPE (in orange), using a correction factor assuming that the signal-to-variance ratio is 0.4, against the true worst-case MPSE (in green). We observe that if the true signal-to-variance ratio is smaller than our assumption of 0.4, then our estimate is too conservative, and vice versa if the true signal-to-variance ratio is larger.

5.3. Causal and anti-causal predictors

We demonstrate the ability of xPAR to select causal predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data from an SCM (Fig. 7 [top]), where one anchor, A_1, is a parent of the causal predictors, while the other, A_2, is a parent of the anti-causal predictors. We consider two

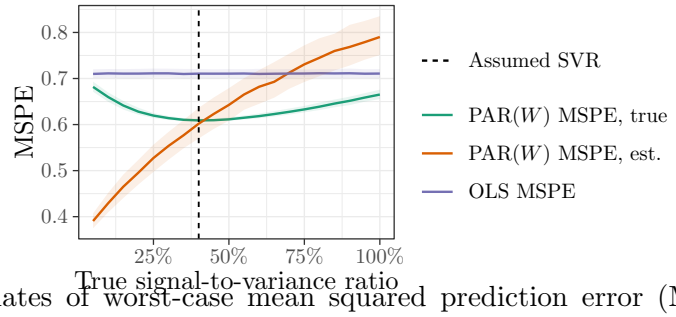


Figure 6.: Estimates of worst-case mean squared prediction error (MSPE) over a robustness set C. PAR is applied assuming that the signal-to-variance ratio is 0.4, which gives an estimate of the worst-case MSPE over C (orange). Green line shows actual worst-case MSPE over C at different underlying signal-to-variance ratios.

identically distributed noisy proxies W, Z of $A := (A_1, A_2)$. The challenge is that A_2 is measured with significantly more noise than A_1, across both proxies.

As seen in Fig. 7 [bottom] PAR places more weight on anti-causal features. In effect, the noise in the measurement of A_2 causes $X_{\text{anti-causal}}$ to appear less sensitive to shifts in A_2. This is an ideal scenario for xPAR, as it is designed to deal with additional noise by leveraging both proxies. Consequently, when two proxies W, Z are available, xPAR places more weight on the causal predictors, relative to PAR.

5.4. Targeted shift

We demonstrate the trade-off made by Targeted Anchor Regression (TAR) versus Anchor Regression (AR), considering the case when A is observed for simplicity. We simulate training data and fit estimators γ_{OLS}, γ_{AR} and γ_{TAR}, where γ_{TAR} is targeted to a particular mean and covariance of a random intervention ν, and we select λ for γ_{AR} such that this intervention is contained within $C_A(\lambda)$.

We then simulate test data from two distributions: $\mathbb{P}_{do(A:=\nu)}$ (i.e., the shift occurs), and $\mathbb{P}$ (where it does not), and evaluate the mean squared prediction error (MSPE). The results are shown in Fig. 8, and demonstrated that TAR performs better than AR and OLS in the first scenario, but this comes at the cost of worse performance on the training distribution.

6. Real-Data Experiment: Pollution

We test our approach on a real-world heterogeneous dataset of hourly pollution readings in five cities in China, taken over several years [Liang et al., 2016], with most data available from 2013-15. Our prediction target is PM2.5 concentration, a measure of pollution, and covariates are primarily weather-related, including dew point, temperature, humidity, pressure, wind direction / speed, and precipitation.

Real-World Proxy (Temperature): Pollution tends to be seasonal in this dataset, and so we construct our training and test environments using seasons: For each of the

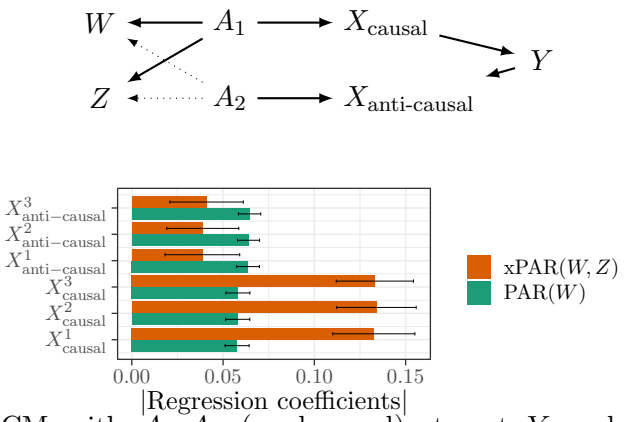


Figure 7.: *Top*: SCM with A_1, A_2 (unobserved), target Y and predictor variables $X_{\text{causal}}, X_{\text{anti-causal}} \in \mathbb{R}^3$. Dotted lines indicate higher noise. *Bottom*: Absolute value of regression coefficients. PAR places more weight on anti-causal predictors, while xPAR places more weight on causal predictors.

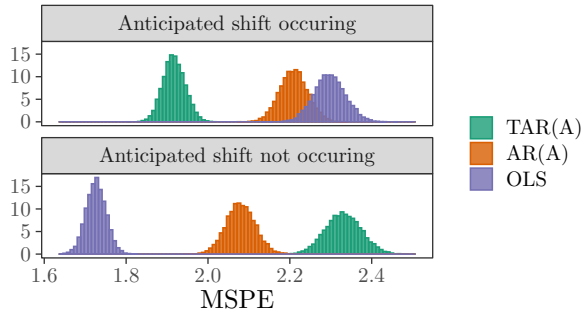


Figure 8.: Empirical mean squared prediction error of TAR, OLS and AR under the shifted distribution and the training distribution.

four seasons, we train only on the other three seasons, and evaluate on the held-out season. We do this for each city, treating each city and held-out season as a separate evaluation. This leads to 20 separate scenarios.

With this variation in mind, we use temperature as a real-world proxy, and treat it as unavailable at test time. We also construct two noisier copies of temperature, which we refer to as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\text{var}(\text{Temp})/\,\text{var}(W) = 0.9$.

Estimators / Benchmarks: For Proxy and Cross-Proxy AR (PAR, xPAR, see Section 3), we choose $\lambda \in [0, 40]$ by leave-one-group-out cross-validation on the three training seasons, using the first year (2013) of data. For instance, if "winter" is the test season, then we choose the value of λ that performs best on average across combinations of the other seasons e.g., training on the fall & summer data and evaluating on the spring data.

When using temperature as a single proxy in PAR, we observe that in 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent

Table 1.: Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
PAR (TempC)	0.531	6	-0.041	0.006
PAR (W)	0.531	6	-0.037	0.006
xPAR (W, Z)	0.531	6	-0.039	0.007
PTAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)	0.529	8	-0.038	0.001
xPTAR (W, Z)	0.526	7	-0.059	0.001

to OLS. For comparability, we use the same values of λ for PAR(W) and xPAR(W, Z). For Proxy Targeted AR and Cross-Proxy Targeted AR (PTAR, xPTAR, see Section 4), we use the mean and variance of the relevant variables (e.g., temperature, W, Z) in the held-out season to target our predictors.

Our primary benchmark is OLS (without temperature). We also compare to (a) OLS that uses temperature during train and test [OLS (TempC)], and (b) OLS that includes the temperature during training, and uses the mean test value for temperature during prediction [OLS + Est. Bias]. We present the results for the 9 scenarios where $\lambda > 0$ in Table 1, since PAR with $\lambda = 0$ is equivalent to OLS (aggregate results in Table C.1 in the supplement).

Results: For both PAR and PTAR, we see improvement over OLS on average across scenarios, with limited downside (e.g., in the worst scenario for PTAR relative to OLS, the additional MSE incurred is 0.001). In Fig. C.4 (Supplement), we observe that PAR and PTAR achieve gains in two different ways: PAR increases the coefficients of humidity and dew point relative to OLS, while PTAR reduces them and incorporates a correction into the intercept.

7. Discussion and Related Work

Learning a predictive model that performs well under arbitrarily strong causal interventions is an ambitious goal. In this work, we have argued that even if causal invariance is achievable, it may not be desirable: A model whose performance is invariant to arbitrarily strong interventions may have poor performance when the test distribution does not differ too much from the training distribution.

There is a large body of work that seeks to learn causal models as a route to achieving

invariance [Rojas-Carulla et al., 2018, Magliacane et al., 2018], or that uses knowledge of the causal graph to select predictors with invariant performance under a set of known interventions [Subbaswamy et al., 2019]. Similarly, invariant risk minimization (IRM) seeks a predictor Φ such that $\mathbb{E}(Y \mid \Phi(X))$ is invariant across a set of discrete environments [Arjovsky et al., 2019, Xie et al., 2020, Krueger et al., 2020, Bellot and van der Schaar, 2020]. Recent work has pointed to the theoretical and practical difficulty of learning such a predictor for IRM [Rosenfeld and Risteski, 2020, Kamath et al., 2021, Guo et al., 2021], in part due to the fact that recovering a truly invariant model, even in linear settings, requires a large number of environments. Generalization in non-linear settings requires sufficient overlap between environments and strong restrictions on the model class [e.g., Christiansen et al., 2021]. In contrast to all of the above, we trade off between in-distribution performance and invariance explicitly, instead of seeking invariance as a primary goal. Moreover, since we allow for A to influence Y directly and through hidden variables, invariance may not even be achievable, but we can still formulate a worst-case loss for bounded interventions.

We argue for incorporating prior knowledge about potential shifts by (1) identifying proxies for relevant factors of variation (i.e., anchor variables), and (2) specifying plausible sets of interventions on these factors of variation. We build upon the causal framework of Anchor Regression [Rothenhäusler et al., 2021], extending it in two important ways.

To start, we relax the assumption that the anchor variables are directly observed. Instead, we only assume access to proxies, and prove that identification of the worst-case loss is feasible with two proxies. The challenge of identifying the worst-case loss is related to the problem of identifying causal effects with noisy proxies of unmeasured confounders [Tchetgen Tchetgen et al., 2020, Miao and Tchetgen, 2018, Shi et al., 2018, Kuroki and Pearl, 2014], and the challenge of learning under classical measurement error [Fuller, 1987, Hyslop and Imbens, 2001, Bound et al., 2001]. Our observation that a single proxy will underestimate the worst-case loss is related to the well-known problem of regression dilution bias [Frost and Thompson, 2000], where performing linear regression under measurement error leads to bias in parameter estimation. In contrast, we are not concerned with causal / structural parameter estimation, which is generally not possible in the models we consider, but rather estimating a worst-case loss under a class of interventions. Srivastava et al. [2020] also consider distributional shift in unmeasured variables for which proxies are available, and apply techniques for handling worst-case sub-populations from DRO [Duchi et al., 2020]. In contrast, we consider causal interventions on A that could lie outside the support of the training data, which cannot be represented as a sub-population. Moreover, they consider the single-proxy case, and give a generalization bound that incorporates the impact of noise, while under our assumptions we are able to recover guarantees as if A were observed, using two proxies.

We then introduce Targeted Anchor Regression, a method for incorporating additional prior knowledge on the strength and direction of shifts in anchor variables. This method can be interpreted as allowing for specification of a broader class of robustness sets, beyond those considered in Rothenhäusler et al. [2021], or as specifying the mean and

covariance of the anchors at test time. We prove analogous results with proxies in this setting, and evaluate this strategy empirically in Section 6, targeting our loss to a particular mean and variance over temperature in the held-out season.

Our work contributes to a growing body of literature that seeks to generalize Anchor Regression to new settings, whether allowing for unobserved anchors and a broader class of robustness sets (as in our work), or generalizing to discrete and censored outcomes, as in Kook et al. [2022].

Acknowledgements

We thank Hussein Mozannar, Chandler Squires, Hunter Lang, Zeshan Hussain, and other members of the ClinicalML lab for feedback and insightful discussions. This work was supported in part by Office of Naval Research Award No. N00014-17-1-2791. NT and JP are supported by a research grant (18968) from VILLUM FONDEN, and JP, in addition, is supported by Carlsberg Foundation.

Invariant Policy Learning: A Causal Perspective

Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister

Abstract

Contextual bandit and reinforcement learning algorithms have been successfully used in various interactive learning systems such as online advertising, recommender systems, and dynamic pricing. However, they have yet to be widely adopted in high-stakes application domains, such as healthcare. One reason may be that existing approaches assume that the underlying mechanisms are static in the sense that they do not change over different environments. In many real-world systems, however, the mechanisms are subject to shifts across environments which may invalidate the static environment assumption. In this paper, we tackle the problem of environmental shifts under the framework of offline contextual bandits. We view the environmental shift problem through the lens of causality and propose multi-environment contextual bandits that allow for changes in the underlying mechanisms. We adopt the concept of invariance from the causality literature and introduce the notion of policy invariance. We argue that policy invariance is only relevant if unobserved variables are present and show that, in that case, an optimal invariant policy is guaranteed to generalize across environments under suitable assumptions. Our results may be a first step towards solving the environmental shift problem. They also establish concrete connections among causality, invariance, and contextual bandits.

1. Introduction

The problem of learning decision-making policies lies at the heart of learning systems. To adopt these learning systems in high-stakes application domains such as personalized medicine or autonomous driving, it is crucial that the learned policies are reliable even in environments that have never been encountered before. In this paper, we consider the problem of learning policies that are robust with respect to shifts across environments. We consider this question in the setup of offline contextual bandits, which provides a mathematical framework for tackling the above learning problems.

While recent studies in offline contextual bandits Dudik et al. [2011], Bottou et al. [2013], Swaminathan and Joachims [2015a,b], Zhou et al. [2018], Kallus [2018], Athey and Wager [2021] offer theoretical results and novel methodologies for policy learning from offline data, they primarily focus on an independent and identically distributed (i.i.d.) setting, in which the underlying mechanisms do not change over time or over different environments. In practice, however, shifts between environments often occur, possibly invalidating the i.i.d. assumption. In healthcare, for example, datasets from different hospitals may not come from the same underlying distribution. As a result, a learning agent that ignores environmental shifts may fail to generalize beyond the environments it was trained on.

In the supervised learning context, the environmental shift problem has been studied under different names, such as domain generalization, covariate shift adaptation, distributional robustness or out-of-distribution generalization Sugiyama and Kawanabe [2012], Muandet et al. [2013], Volpi et al. [2018], Arjovsky et al. [2019], Christiansen et al. [2021]. In domain generalization, the goal is to develop learning algorithms that are robust to changes in the test distribution. Thus, a fundamental problem is how to characterize such changes. A promising direction relies on a causal framework to describe the changes through the concept of interventions Schölkopf et al. [2012], Rojas-Carulla et al. [2018], Magliacane et al. [2018], Arjovsky et al. [2019], Christiansen et al. [2021]. A key insight is that while purely predictive methods perform best if test and training distributions coincide, causal models generalize to arbitrarily strong interventions on the covariates because of the modularity property of structural causal models (see e.g., Pearl [2009]).

The environmental shift problem is related to the problem of transportability in causal inference Pearl and Bareinboim [2011], Bareinboim and Pearl [2014, 2016], Subbaswamy et al. [2019], Lee et al. [2020], Correa and Bareinboim [2020] which aims to generalize causal findings from source environments to a target environment. Unlike our work, transportability assumes knowledge of how the target environment differs from the source environments as well as the underlying causal graph that is shared among them through selection diagrams Pearl and Bareinboim [2011]. Using this causal knowledge, the task of transportability is to derive whether and how one can identify a causal quantity (e.g. an interventional distribution) in the target environment using data obtained from the source environments.

In real-world applications, however, knowledge of the underlying causal graph and structural discrepancies between environments may not be available. In recent years, invariance-based methods have been exploited to learn the causal structure from data Peters et al. [2016], Pfister et al. [2019b], Heinze-Deml et al. [2018]. In invariant causal prediction Peters et al. [2016], for example, one assumes that the data are collected from different environments, each of which describes different underlying data-generating mechanisms, and uses this heterogeneity to learn the causal parents of an outcome variable Y. The underpinning assumption is the invariance assumption, which posits the existence of a set of covariates X in which the mechanism between X and Y remains constant. A model based on such invariant covariates is guaranteed to generalize to all unseen environments.

Although some recent studies have explored the use of causality and invariance for tackling the environmental shift problem in contextual bandits and, more generally, reinforcement learning Zhang et al. [2020], Sonar et al. [2021], the actual roles and benefits of causality and invariance remain unclear and under-explored. Graphical models have also been used in reinforcement learning to represent the underlying Markov Decision Processes (MDP) under the framework of factored MDPs. Such methods, however, focus mainly on providing efficient planning algorithms rather than generalizing to a new environment Kearns and Koller [1999], Guestrin et al. [2003, 2002], Jonsson and Barto [2006].

Our paper delineates an explicit connection among causality, invariance, and the environmental shift problem in the context of contextual bandits. We develop a causal framework for characterizing the environmental shift problem in contextual bandits, and provide a practical and theoretically sound solution based on the proposed framework. Our framework differs from the framework of causal bandits Lee and Bareinboim [2018], Lattimore et al. [2016], Yabe et al. [2018], de Kroon et al. [2020]. While causal bandits exploit causal knowledge (either assumed to be known a priori or estimated from data) for improving the finite sample performance in a single environment, our framework focuses on modeling distributional shifts and the ability to generalize to new environments. Another line of work has addressed the problem of policy evaluation and learning under unobserved confounding between the action and the reward variables Bareinboim et al. [2015], Sen et al. [2017], Tennenholtz et al. [2020], Kallus and Zhou [2020], Tennenholtz et al. [2021]. In contrast, we consider the complementary problem of unobserved confounding between the covariates and the reward variables (see Section 3).

Our contributions are fourfold. First, we propose a multi-environment contextual bandit framework that represents mechanisms underlying a contextual bandit problem by structural causal models (SCMs; Pearl [2009]). The framework allows for changes in environments and thereby relaxes the i.i.d. assumption. We define environments as different perturbations on the underlying SCM, and we evaluate the policy according to its worst-case performance in all environments. Second, using the proposed framework, we generalize the invariance assumption used in methods such as invariant causal prediction and define invariance properties for policies that, under certain assumptions, guarantee generalizability to unseen environments. Third, we develop an offline method for testing invariance under distributional (policy) shifts, and provide an algorithm for finding an optimal invariant policy. Fourth, we highlight a setting in which causality and invariance are not necessary for solving the environmental shift problem. This insight takes us closer to understanding what causality can offer in contextual bandits.

The remainder of our paper is organized as follows. The rest of this section briefly reviews an offline contextual bandit problem. Section 2 formally defines a causal framework for multi-environment contextual bandits and highlights the roles of causality and invariance in this formulation. We show that in the absence of unobserved confounders, causality does not improve the generalization ability in that they do not outperform purely predictive approaches. Drawing on the proposed framework, Section 3 introduces invariance properties for policies and provides the main theoretical contributions underpinning our solution for the environmental shift problem. Section 4 discusses the

assumptions required to learn invariant policies from offline data and presents an algorithm for learning an optimal invariant policy. Section 5 provides simulation experiments that empirically verify our theoretical results. In Section 6, we apply our framework to a warfarin dosing study.

1.1. Offline contextual bandits

We briefly review the offline contextual bandit problem Beygelzimer and Langford [2009], Strehl et al. [2010], considering a setup in which part of covariates (also known as context variables) are unobserved. More precisely, we assume that the covariates can be partitioned into observed and unobserved variables $X \in \boldsymbol{\mathcal{X}}$ and $U \in \boldsymbol{\mathcal{U}}$. Here, $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{U}}$ are metric spaces; the reader may think of $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^d$ and $\boldsymbol{\mathcal{U}} \subseteq \mathbb{R}^p$. As in the standard contextual bandit setup Langford and Zhang [2008], for each round, we assume that the system generates a covariate vector (X, U) and reveals only the observable X to an agent. From the observed covariates X, the agent selects an action $A \in \mathcal{A}$ according to a policy $\pi : \boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})$ that maps the observed covariates to the probability simplex $\Delta(\mathcal{A})$ over the set of actions $\mathcal{A}$. Adapting commonly used notation, we write, for all $x \in \boldsymbol{\mathcal{X}}$ and $a \in \mathcal{A}$, $\pi(a|x) := \pi(x)(a)$. The agent then receives a reward R depending on the chosen action A, and on both the observed and unobserved covariates (X, U).

In the classical setting, one assumes that the covariates are drawn i.i.d. from a joint distribution $\mathbb{P}_{X,U}$ (an assumption we will relax when introducing multi-environment contextual bandits in Section 2) and that the rewards are drawn from a conditional distribution $\mathbb{P}_{R|X,U,A}$. The agent is evaluated based on the performance of its policy π which is measured by the policy value:

$$V(\pi) := \mathbb{E}_{(X,U)\sim\mathbb{P}_{X,U}} \mathbb{E}_{A\sim\pi(X)} \mathbb{E}_{R\sim\mathbb{P}_{R|X,U,A}} \big[R\big].$$

The agent is now given a fixed training dataset that is collected offline: it consists of n rounds from one or more different policies, i.e., $D := \{(X_i, A_i, R_i, \pi_i(X_i))\}_{i=1}^n$, where $A_i \sim \pi_i(X_i)$[5] for all $i \in \{1, \ldots, n\}$. The goal of the agent is then to find a policy π that maximizes the policy value over a given policy class Π, i.e., $\pi^* \in \arg\max_{\pi\in\Pi} V(\pi)$.

This setting assumes that the covariates in each round are sampled i.i.d. from some fixed distribution; this implies that the environment in which we deploy the agent is identical to the environment in which the training dataset was collected. Section 2 introduces a causal framework for multi-environment contextual bandits, a framework that relaxes the i.i.d. assumption.

[5] We assume knowledge of the initial policy π_i to ease our presentation and focus our contribution on the environmental shifts problem. Our theoretical results and algorithms remain unchanged even if the initial policy is unknown and needs to be estimated from the offline data (see Appendix D.5 for more details).

2. A Causal Framework for Multi-environment Contextual Bandits

Instead of having a fixed distribution $\mathbb{P}_{X,U}$ over the covariates, we introduce a collection $\mathcal{E}$ of environments such that, in each round, the covariates are drawn from an environment-specific distribution $\mathbb{P}^e_{X,U}$ that depends on the environment $e \in \mathcal{E}$ in that round.

In practice, the agent only observes part of the environments $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}$ and is expected to generalize well to all environments in $\mathcal{E}$ including the unseen environments $\mathcal{E} \setminus \mathcal{E}^{\text{obs}}$. To formalize the problem, we first introduce a framework that puts assumptions on how environments change the distributions of X, U and R. Specifically, an environment e can only perturb the distribution of the reward R through altering the distribution of the observed covariates X. This constraint makes it possible to generalize information learned from one set of environments to another. In this formulation – even though the full conditional distribution of the reward $\mathbb{P}^{\pi,e}_{R|X,U,A}$ is assumed to be fixed across environments – the observable distribution $\mathbb{P}^{\pi,e}_{R|X,A}$ after marginalizing out the unobserved U may change from one environment to another (see, e.g., Fig. 1b)

Formally, the assumptions are constructed via an underlying class of SCMs indexed by the environment and policy.[6]

Setting 1 (Multi-environment Contextual Bandits)**.** *Let $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \ldots \times \mathcal{X}^d$ and $\boldsymbol{\mathcal{U}} = \mathcal{U}^1 \times \ldots \times \mathcal{U}^p$ be products of metric spaces, $\mathcal{A} = \{a^1, \ldots, a^k\}$ a discrete action space, $\Pi := \{\boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})\}$ the set of all policies, and $\mathcal{E}$ a collection of environments. For all $\pi \in \Pi$ and all $e \in \mathcal{E}$ we consider the following SCMs,*

$$\mathcal{S}(\pi, e): \quad \begin{cases} U \coloneqq s(X, U, \varepsilon_U) \\ X \coloneqq h_e(X, U, \varepsilon_X) \\ A \coloneqq g_\pi(X, \varepsilon_A) \\ R \coloneqq f(X, U, A, \varepsilon_R), \end{cases} \tag{1}$$

where $(X, U, A, R) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{U}} \times \mathcal{A} \times \mathbb{R}$, s, $(h_e)_{e\in\mathcal{E}}$, and f are measurable functions, $\varepsilon = (\varepsilon_U, \varepsilon_X, \varepsilon_A, \varepsilon_R)$ is a random vector with independent components and a distribution $Q_\varepsilon = Q_{\varepsilon_U} \otimes Q_{\varepsilon_X} \otimes Q_{\varepsilon_A} \otimes Q_{\varepsilon_R}$, and g_π and Q_{ε_A} are such that for all $x \in \boldsymbol{\mathcal{X}}$ it holds that $g_\pi(x, \varepsilon_A)$ is a random variable on $\mathcal{A}$ with distribution $\pi(x)$. Fig. 1a visualizes the coarse-grained structure of this setting. U, X, and A should be thought of as random vectors. Accordingly, h_e, for example, is a function with a multivariate output; it is a short-hand notation in the sense that a component of h_e does not need to depend on all X, for example. In particular, we assume that the graph $\mathcal{G}$ (defined below) corresponding to the SCMs is acyclic, see Fig. 1b and 1c for an example.

[6] Readers familiar with the standard notion of SCMs may think about an SCM with a source node E. $\mathcal{S}(\pi, e)$ then corresponds to an intervention on the action variable (change of policy) and on some of the observed covariates variables (change of environment). Here, we consider fixed environments, so that we do not have to consider them as random draws from an underlying distribution; see also Dawid [2002].

We assume there exists a probability measure μ on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{U}} \times \mathcal{A} \times \mathbb{R}$ such that for all $\pi \in \Pi$ and all $e \in \mathcal{E}$ the SCM $\mathcal{S}(\pi, e)$ induces a unique distribution $\mathbb{P}^{\pi,e}$ over (X, U, A, R) (see Bongers et al. [2021] for details) which is dominated by μ and has full support on X. We denote the corresponding density by $p^{\pi,e}$ and the corresponding expectations by $\mathbb{E}^{\pi,e}$. Whenever a probability, density, or expectation does not depend on π, we omit π and write $\mathbb{E}^e[X]$ rather than $\mathbb{E}^{\pi,e}[X]$, for example.

Some remarks regarding Setting 1 are in order: (1) We only use the SCMs as a flexible way of modeling the changes in the joint distribution with respect to the environment e and the policy π. In particular, we do not use it to model any further intervention distributions that do not correspond to a change of policy or environment. (2) In practice, the precise form of the SCMs is unknown. Indeed, we will neither assume knowledge of the structural equations nor complete knowledge of the graph structure, except that the constraints induced by (1) hold. (3) The assumption of a dominating measure for all environments ensures that we can always assume the existence of densities while also switching across environments. In particular, this avoids any measure-theoretic difficulties regarding conditional distributions. (4) The assumption that the induced distributions over X have full support in all environments ensures that the generalization problem when moving from $\mathcal{E}^{\text{obs}}$ to $\mathcal{E}$ does not involve any extrapolation. Additionally, it ensures that conditional expectations such as $\mathbb{E}^{\pi,e}[R \mid X = x]$ can be uniquely defined for all $x \in \boldsymbol{\mathcal{X}}$ as integrals of the conditional densities. (5) The environments are modelled fixed (and not random). However, we could also treat the environments as random variables which can be considered a special case of the fixed-environment setting (see Appendix D.4). (6) The assumption that U is not affected by the environments is necessary for the existence of a d-invariant set (Definition 3). If the assumption is violated, there is no d-invariant set.

We now introduce the graph $\mathcal{G}$ over the variables $(X^1, \dots, X^d, U^1, \dots, U^p, A, R)$ that visualizes the structure of the SCMs $\mathcal{S}(\pi, e)$ (for all $\pi \in \Pi$ and $e \in \mathcal{E}$). More precisely, $\mathcal{G}$ is constructed as follows: Each coordinate of the variables (X, U, A, R) corresponds to a node. The nodes are connected according to the structural assignments, that is, we draw a directed edge from a variable B to a variable C if, for at least one environment $e \in \mathcal{E}$, the variable B appears on the right-hand side of the structural assignment of variable C (see Fig. 1b for an example). Let $\mathcal{I} \subseteq \{1, \dots, d\}$ index the variables X^j for which the structural assignment $X^j := h_e^j(X, U, \varepsilon_X)$ in (1) varies with e, i.e., where there exist $e, f \in \mathcal{E}$ such that $h_e^j \neq h_f^j$. The environments $\mathcal{E}$ correspond to perturbations on variables $X^{\mathcal{I}}$, which implies that for each $e \in \mathcal{E}$ the distribution $\mathbb{P}^{\pi,e}(X^{\mathcal{I}} \mid U, X^{\{1,\dots,d\}\setminus\mathcal{I}})$ may vary. We augment the graph with a square node e to represent the environments and draw a directed edge from the node e to each of the perturbation targets $X^{\mathcal{I}}$. Furthermore, we draw edges from all nodes in X to A and mark them with π (to represent their dependence on the policy).

This graph $\mathcal{G}$ is assumed to be acyclic, that is, to not contain any directed cycles. Some of the theoretical results require an additional assumption on the structure of the graph $\mathcal{G}$: there is no unobserved variable U that influences only the observed X but not the reward R, see Assumption 1 below.

Assumption 1. *Let $\mathcal{G}$ be the graph of the SCMs in Setting 1. We assume that, for all $\ell \in \{1, \dots, p\}$, there must be an edge from U^ℓ to R in $\mathcal{G}$.*

By the Markov condition, which holds in SCMs, the graph $\mathcal{G}$ defined above encodes (conditional) independence statements, which we will see relate to invariance, through the concept of d-separation. More precisely, the Markov condition states that any d-separation statement in a graph implies conditional independence Pearl [2009], Lauritzen et al. [1990], Peters et al. [2017]. Here, we refer to the standard definition of d-separation when not distinguishing between the different types of nodes and denote by $\perp\!\!\!\perp_{\mathcal{G}}$ a d-separation statement in a graph $\mathcal{G}$. For completeness, we define d-separation in Appendix D.1. In this work, however, the distribution and conditional dependencies depend on the policy π, which motivates to consider graphs that change accordingly. For any $S \subseteq \{1, \dots, d\}$, we therefore define $\mathcal{G}^S$ to be the graph that is equal to $\mathcal{G}$ but only has edges from X^S to A (rather than from all X to A). For any $\pi \in \Pi^S$, the distribution is then Markov with respect to $\mathcal{G}^S$, see Lemma D.1 in Appendix D.3.3.

We are now ready to define contextual bandits with multiple environments.

Definition 1 (Multi-environment Contextual Bandits)**.** Assume Setting 1. In a multi-environment contextual bandit setup, before the beginning of each round, the system is in an environment $e \in \mathcal{E}$ Then, the system generates a covariate vector (X, U) and reveals only the observable X and the environment label e to the agent. Based on the observed covariates X, the agent selects an action A according to the policy $\pi : \boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})$. The agent then receives a reward R, depending on the chosen action A and on both the observed and unobserved covariates (X, U). More precisely, we assume for all $i \in \{1, \dots, n\}$ that (X_i, U_i, A_i, R_i) are sampled independently according to $\mathbb{P}^{\pi_i, e_i}_{X,U,A,R}$ (see Setting 1). The training data contains data from environments in $\mathcal{E}^{\text{obs}}$. When $\left|\mathcal{E}^{obs}\right| = |\mathcal{E}| = 1$, the setup reduces to a standard contextual bandit setup.

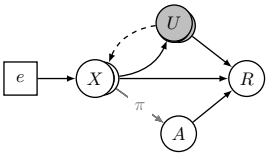
In the multi-environment contextual bandit setup, the covariates on different rounds are not identically distributed due to changes in the environments. We can thus use this framework to model situations, where the test environments differ from training environments. We illustrate this setting with the following example, which we will refer back to several times throughout the paper.

Example 1 (Linear Confounded Multi-environment Contextual Bandits)**.** *Consider a linear multi-environment contextual bandit with the following underlying SCMs*

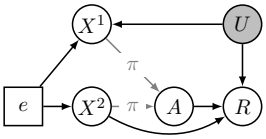
$$\mathcal{S}(\pi, e): \quad \begin{cases} U \coloneqq \varepsilon_U \\ X^1 \coloneqq \gamma_e U + \varepsilon_{X^1} \\ X^2 \coloneqq \alpha_e + \varepsilon_{X^2} \\ A \coloneqq g_\pi(X^1, X^2, \varepsilon_A) \\ R \coloneqq \begin{cases} \beta_1 X^2 + U + \varepsilon_R, & \text{if } A = 0 \\ \beta_2 X^2 - U + \varepsilon_R, & \text{if } A = 1, \end{cases} \end{cases}$$

where $\varepsilon_R, \varepsilon_A, \varepsilon_{X^1}, \varepsilon_{X^2}$ are jointly independent noise variables with zero mean, $\gamma_e, \alpha_e \in \mathbb{R}$ for all $e \in \mathcal{E}$, $\beta_1, \beta_2 \in \mathbb{R}$, and $\mathcal{A} = \{0, 1\}$. Fig. 1b depicts the induced graph $\mathcal{G}$. In this example, the environments influence the observed covariates in two ways: (a) they change the mean of X^2 via α_e and (b) they change the conditional mean of X^1 given U via γ_e, while the rest of the components remain fixed across different environments. Here, the environment-specific coefficient γ_e modifies the correlation between the observable X^1 and the unobserved variable U, and consequently between X^1 and the reward R. Thus, an agent that uses information from X^1 to predict the reward R in the training environments may fail to generalize to other environments. To see this, consider a training environment $e = 1$ and a test environment $e = 2$ and let $\gamma_1 = 1$, $\gamma_2 = -1$ be the environment-specific coefficients in the training and test environment, respectively. In the training environment, we have a large positive correlation between X^1 and U, and consequently the agent will learn that the action $A = 0$ yields a higher expected reward when observing a positive value of X^1 (and $A = 1$ otherwise). However, the correlation between X^1 and U becomes negative (and large in absolute value) in the test environment, which means that the policy that the agent learned from the training environment will now be harmful. We will see in Section 3 that a policy that depends on a d-invariant set ($\{X^2\}$ in this example) does not suffer from this generalization problem and is guaranteed to generalize across different environments.

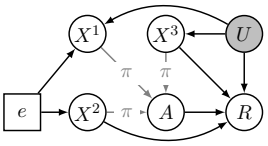
A similar structure appears in the medical example discussed in Section 6. There, A is the dose of a drug, R is a response variable, X are observed patient features and U are unobserved genetic factors. The environment e is (a proxy of) the continent on which the data was collected.



(a) A summary of the causal structure of Setting 1 (with dashed edge) or Setting 2 (without dashed edge). In both cases the causal relations between X- and U-covariates are not shown explicitly.



(b) Graph $\mathcal{G}$ induced by the data-generating process in Example 1. $\mathcal{G}^{\{2\}}$ is the graph obtained when removing the edge $X^1 \to A$.



(c) Graph $\mathcal{G}$ induced by a more complicated data-generating process than in (b).

Figure 1.: Graphs summarizing different data-generating models. White and grey circles represent observed and hidden variables, respectively. (a) A summary graph depicting the causal structure. If the dashed edge from U to X is included it corresponds to Setting 1 otherwise to Setting 2. (b) The graph induced by the model in Example 1. Here, $\{X^2\}$ is d-invariant, because $R \perp\!\!\!\perp_{\mathcal{G}^{\{2\}}} e \mid X^S$, see Definition 3. Any set S that contains X^1 is not d-invariant because of the open path $e \to X^1 \leftarrow U \to R$. In this example, a policy depending on X^1 and X^2 may perform worse in a test environment than (a d-invariant) policy depending only on X^2. In practice, we do not assume that the structure is known but test for invariances (11) from data. This requires testing under distributional shifts: even though $\{X^2\}$ is d-invariant, (11) may not hold for a policy π that depends on X^1 and X^2 because of the path $e \to X^1 \to A \to R$. (c) A more complex model, where the environments do not act on all X variables. Although U has an edge into X^3, the subset $\{X^2, X^3\}$ is still a d-invariant set – there is no edge from e to X^3. Again, every subset of variables containing X^1 is not d-invariant. (In fact, in examples (b) and (c), X^1 is a strongly non-d-invariant variable, see Definition 5, and cannot be part of a d-invariant set.)

2.1. Distributionally robust policies

To evaluate the performance of an agent across different environments, we define a policy value that takes into account environments. In particular, we focus on the worst-case performance of an agent across environments.

Definition 2 (Robust Policy Value)**.** For a fixed policy $\pi \in \Pi$, and a set of environments $\mathcal{E}$, we define the *robust policy value* $V^{\mathcal{E}}(\pi) \in \mathbb{R}$ as the worst-case expected reward

$$V^{\mathcal{E}}(\pi) \coloneqq \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi, e}\left[R\right]. \tag{2}$$

Intuitively, an agent that maximizes the robust policy value is expected to perform well (relative to other agents) in the most harmful environment. The idea of optimizing worst-case performance has been suggested in the reinforcement learning literature García and Fernández [2015], Amodei et al. [2016] to ensure safe behavior of an agent and prevent catastrophic events and has also been used to formulate adversarial training Bai et al. [2021] as well as out-of-distribution generalization Ye et al. [2021].

We now assume that, for several observed environments, we are given an i.i.d. sample from a multi-environment contextual bandit, see Definition 1. More precisely, we assume to observe $D \coloneqq \{(X_i, A_i, R_i, \pi_i(X_i), e_i)\}_{i=1}^n$, where $e_i \in \mathcal{E}^{\text{obs}}$, $A_i \sim \pi_i(X_i)$, $(X_i, A_i, R_i) \overset{\text{ind.}}{\sim} \mathbb{P}_{X,A,R}^{\pi_i, e_i}$ for all $i \in \{1, \dots, n\}$. Using only D, we aim to solve the following maximin problem[7]:

$$\arg\max_{\pi \in \Pi} V^{\mathcal{E}}(\pi). \tag{3}$$

If we do not observe all the environments, solving the maximin problem (3) is impossible without further assumptions. A baseline approach to this problem is to pool the data from all training environments and learn a policy that maximizes the policy value ignoring the environment structure. We will see in Section 2.2 that this is indeed optimal if all relevant covariates have been observed. If, however, hidden confounding is present, the pooling approach does not necessarily yield an optimal policy and the learned policy may fail to generalize to unseen test environments. In Section 3, we introduce the notion of policy d-invariance. We will show that under certain assumptions, solving the maximin problem (3) amounts to finding an optimal d-invariant policy which is then guaranteed to generalize across environments.

2.2. Policy learning without unobserved confounders

This section illustrates a setting in which it is not beneficial to explicitly take into account the environment structure. Here, simply pooling the data from all training environments and applying a standard value-based policy learning algorithm yields a solution to (3). This result sheds light on the role of causality and invariance in contextual bandits and reinforcement learning. The following setting is a modification of Setting 1 without any unobserved confounders in the causal model.

[7] The maximum can always be attained when Π is an unrestricted policy class and takes a form similar to (4).

Setting 2 (Unconfounded Multi-environment Contextual Bandits)**.** *Assume the same setup as in Setting 1 but assume that the SCMs are given by*

$$\mathcal{S}(\pi, e): \quad \begin{cases} U \coloneqq s(X, U, \varepsilon_U) \\ X \coloneqq h_e(X, \varepsilon_X) \\ A \coloneqq g_\pi(X, \varepsilon_A) \\ R \coloneqq f(X, U, A, \varepsilon_R). \end{cases}$$

Fig. 1a (with the dashed line from U to X excluded) summarizes the causal structure of Setting 2.

The following theorem shows that in Setting 2 there is a population optimal policy that does not depend on the environments. In particular, this optimal policy can be learned from data obtained in any observed subset of the environments $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}$.

Theorem 1. *Assume Setting 2, let $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}$ be a non-empty subset of observed environments. Let $\pi^* \in \Pi$ be a policy such that for all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$*

$$\pi^*(a|x) > 0 \implies a \in \arg\max_{a' \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x, a'), \tag{4}$$

where $Q^{\mathcal{E}^{\text{obs}}}(x, a) \coloneqq \frac{1}{|\mathcal{E}^{\text{obs}}|} \sum_{e \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi_a, e}[R \mid X = x]$ and π_a is the policy that always selects a [8]. Then,

$$\pi^* \in \arg\max_{\pi \in \Pi} V^{\mathcal{E}}(\pi),$$

i.e., π^* is a solution to the maximin problem (3).

Proof. See Appendix D.3.1. □

This type of generalization is well-established in the context of regression. In the contextual bandit setting the value function $\mathbb{E}^{\pi, e}[R]$ changes across environments, so instead one needs to use that the Q-function $Q^e(x, a) = \mathbb{E}^{\pi_a, e}[R \mid X = x]$ does not change across environments $e \in \mathcal{E}$ and then argue that this implies that the optimal policy remains the same in each environment. Theorem 1 suggests that we can estimate an optimal policy by pooling the data from training environments and applying a standard value-based policy learning algorithm. This is indeed the case, see Appendix D.2 for the consistency result.

Theorem 1 implies that on the population level without hidden confounders, we do not benefit from taking into account the environment structure. However, the following section shows that this is different when hidden confounders exist.

[8] The conditional expectation under π_a can also be written in terms of do-notationPearl [2009], that is, $\forall a \in \mathcal{A}, x \in \mathcal{X} : \mathbb{E}^{\pi_a, e}[R \mid X = x] = \mathbb{E}^e[R \mid X = x, do(A = a)]$. We use the π_a notation to make our presentation consistent.

3. Invariant Policies for Distributional Robustness

We now consider the setting with hidden confounders (Setting 1). This section introduces d-invariant sets and policies, and shows that, under Setting 1, the maximin problem (3) can be reduced to finding an optimal d-invariant policy given certain assumptions, see Proposition 1 and Theorem 2.

We first define d-invariant sets (d for d-separation)[9].

Definition 3 (d-invariant Sets)**.** A subset $S \subseteq \{1, \dots, d\}$ is said to be *d-invariant* if the following d-separation statement holds:

$$R \perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S. \tag{5}$$

Next, we define d-invariant policies. For all subsets $S \subseteq \{1, \dots, d\}$, let us denote the set of all policies that depend only on X^S by $\Pi^S := \{\pi \in \Pi \mid \exists \pi^S : \boldsymbol{\mathcal{X}}^S \to \Delta(\mathcal{A}) \text{ s.t. } \forall x \in \mathcal{X}\,, \pi(\cdot|x) = \pi^S(\cdot|x^S)\}$.

Definition 4 (d-invariant Policies)**.** A policy π is said to be *d-invariant with respect to a subset* $S \subseteq \{1, \dots, d\}$ if S is a d-invariant set and $\pi \in \Pi^S$.

We denote by $\mathbf{S}_{\text{inv}} := \{S \subseteq \{1, \dots, d\} \mid S \text{ is } d\text{-invariant}\}$ the collection of all d-invariant sets and $\Pi_{\text{inv}} := \{\pi \in \Pi \,|\, \exists S \text{ s.t. } \pi \text{ is } d\text{-invariant w.r.t. } S\}$ the collection of d-invariant policies.

d-invariant sets and policies play a central role in solving the distributionally robust objective (3) as illustrated in Proposition 1 and Theorem 2 below. We recall that only part of the environments $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}$ are observed. For now, we assume to have access to the set of d-invariant policies Π^{inv}. Section 4 discusses when and how we can learn Π^{inv} from the observed data.

Because of the hidden confounding, the conditional mean $\mathbb{E}^{\pi,e}[R \mid X = x]$ is, unlike in Section 2.2, not ensured to be stable over the environments. Nevertheless, a d-invariant policy ensures that parts of the conditional mean are unchanged across environments, as shown in the lemma below.

Lemma 1. *Let S^{inv} be a d-invariant set and $\pi^{\text{inv}} \in \Pi^{S^{\text{inv}}}$. It holds for all $e, f \in \mathcal{E}$ and $x \in \boldsymbol{\mathcal{X}}^{S^{\text{inv}}}$ that*

$$\mathbb{E}^{\pi^{\text{inv}},e}\left[R \mid X^{S^{\text{inv}}} = x\right] = \mathbb{E}^{\pi^{\text{inv}},f}\left[R \mid X^{S^{\text{inv}}} = x\right]. \tag{6}$$

Proof. See Appendix D.3.3. □

This implies the following proposition.

[9] We use the term 'd-invariant' to emphasize that the definition is based on the d-separation statement (6) and involves the unseen environments. In related contexts, sometimes the term 'generalizing' is used Pfister et al. [2019a]. Section 4 introduces the invariance hypothesis (11) that is testable from the observed data and discusses the assumptions required to connect the two conditions.

Proposition 1. *Assume Setting 1 and Assumption 1, and that* Π_{inv} *is non-empty. Consider an optimal d-invariant policy* π^* *that maximizes the pooled policy value under the observed environments, that is,* $\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \sum_{e \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi,e}[R]$*. It then holds that*

$$\forall \pi \in \Pi_{\text{inv}} : \qquad V^{\mathcal{E}}(\pi) \leq V^{\mathcal{E}}(\pi^*). \tag{7}$$

Proof. See Appendix D.3.5. □

The key argument in the proof of Proposition 1 is the identifiability of the optimal d-invariant set. Assumption 1 is necessary for this identifiablity: if the assumption is violated and there are multiple d-invariant sets, one can, in general, not say which of those d-invariant sets is optimal with respect to all environments $\mathcal{E}$ (see Appendix D.11 for a more detailed discussion). While, without Assumption 1, the d-invariant set that is most predictive on $\mathcal{E}^{\text{obs}}$ is no longer guaranteed to be worst-case optimal, it still satisfies a weaker guarantee shown in Item 2(i) below.

Proposition 1 shows that a d-invariant policy that is optimal under the observed environments outperforms all other d-invariant policies, even on the test environments. But what about other policies that are not d-invariant? We will see in Theorem 2 that under certain assumptions on the set $\mathcal{E}$ of environments, they cannot perform better than the above π^* either.

We now outline the assumptions on the set $\mathcal{E}$ of environments facilitating this result. As seen in Example 1, the crucial difference between a d-invariant policy $\pi^{\{2\}}$ (a policy that only depends on X^2) and a non-d-invariant policy $\pi^{\{1,2\}}$ (a policy that depends on both X^1 and X^2) is that $\pi^{\{1,2\}}$ can use information related to variables confounded with the reward (X^1 in this example) that may change across environments. In cases where the environments do not change the system 'too strongly' it can therefore happen that using such information is beneficial across all environments. In practice, however, one might not know how strong the test environments can change the system in which case such information can become useless or even harmful. Intuitively, this happens, for example, if environments exist where the non-d-invariant confounded variables no longer contain any information about the reward. Formally, we make the following definition.

Definition 5 (Confounding Removing Environments)**.** For $j \in \{1, \dots, d\}$, we say that the variable X^j is strongly non-d-invariant if for all $S \subseteq \{1, \dots, d\}$

$$R \not\perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^{S \cup \{j\}}.$$

An environment $e \in \mathcal{E}$ is said to be a confounding removing environment if for all $\pi \in \Pi$ it holds that

$$X^j \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} U, \tag{8}$$

for all strongly non-d-invariant variables X^j, where $\mathcal{G}^{\pi,e}$ is the graph induced by the SCM $\mathcal{S}(\pi, e)$.

The two d-separation statements in Definition 5 are in different graphs: Both graphs

$\mathcal{G}^S$ and $\mathcal{G}^{\pi,e}$ are subgraphs of $\mathcal{G}$. The distinction that is important for this definition is that while $\mathcal{G}^S$ contains all edges between the covariates (X, U) that appear in at least one environment, the graph $\mathcal{G}^{\pi,e}$ only contains the edges that are active in the environment $e \in \mathcal{E}$. Furthermore, to provide more understanding of the strongly non-d-invariant variables, we characterize a graphical criterion for such variables in Appendix D.3.4. There we show that the strongly non-d-invariant variables are the variables that are directly affected by e and are confounded with R through U, and descendants of such variables. These strongly non-d-invariant variables should not be included if one wants to find d-invariant sets. For example in Fig. 1c, the variable X^1 is strongly non-d-invariant and the d-invariant sets $\{X^2\}$ and $\{X^2, X^3\}$ are the sets that do not contain X^1.

To give an example of a confounding removing environment, consider the graph $\mathcal{G}^S$ in Example 1 (see Fig. 1b). For any subset S where $\{1\} \subseteq S$ the path $e \to X^1 \leftarrow U \to R$ is open, and therefore X^1 is strongly non-d-invariant. A confounding removing environment is an environment that removes the incoming edge from U to X^1. In such an environment, the variable X^1 does not contain any information about the reward R. A similar notion of confounding removing environments is used in Christiansen et al. [2021] in the setting of prediction.

The existence of confounding removing environments implies that at least in some of the environments it is impossible to benefit from a non-d-invariant policy. To ensure that one cannot benefit in the worst-case, we therefore introduce the following assumption.

Assumption 2 (Strong Environments)**.** *For all $e \in \mathcal{E}$, there exists $f \in \mathcal{E}$ such that f is a confounding removing environment and it holds that $\mathbb{P}^e_X = \mathbb{P}^f_X$.*

To give an example, let $\mathcal{I} \subseteq \{1, \dots, d\}$ index the variables X^j for which there is an edge from e to X^j in the graph $\mathcal{G}$. If the set $\mathcal{E}$ of environments consists of arbitrary interventions on $X^{\mathcal{I}}$, then Assumption 2 is satisfied.

Theorem 2. *Assume Setting 1 and that Π_{inv} is non-empty. Let π^* be an optimal d-invariant policy that maximizes the pooled policy value under the observed environments, $\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \sum_{e \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi,e}[R]$. We then have the following statements.*

(i) We have for all $e \in \mathcal{E}$ that

$$\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a,e}[R] \leq \mathbb{E}^{\pi^*,e}[R]. \tag{9}$$

(ii) If Assumptions 1 and 2 hold, we have

$$\forall \pi \in \Pi: \qquad V^{\mathcal{E}}(\pi) \leq V^{\mathcal{E}}(\pi^*). \tag{10}$$

Proof. See Appendix D.3.6. □

The first statement of Theorem 2 implies that in all environments the expected reward under an optimal d-invariant policy is larger than any optimal context-free policy. In other words, the information gained from the d-invariant set of covariates (the set that

π^* depends on) is generalizable across environments in the sense that it is not harmful in any environment. The second statement states that if the environments $\mathcal{E}$ are sufficiently strong (Assumption 2) then an optimal d-invariant policy π^* maximizes the robust policy value $V^{\mathcal{E}}$.

The above results motivate a procedure to solve the distributionally robust objective (3). Proposition 1 implies that if we consider a policy class containing only the d-invariant policies, the maximin problem reduces to a standard policy optimization problem. Theorem 2 shows that an optimal d-invariant policy, under the assumption of strong environments, is a solution to the distributionally robust objective. In other words, given a training dataset D, we seek to operationalize the following two steps: (a) find the set Π_{inv} of all d-invariant policies (Section 4.1 discusses under which assumptions this is possible), (b) use offline policy optimization to solve $\arg\max_{\pi \in \Pi_{\text{inv}}} V^{\mathcal{E}^{\text{obs}}}(\pi)$ on the data set D.

One of the key components of the proposed method is to test whether a policy π, which may be different from the policy generating the data, is d-invariant using data obtained from the observed environments $\mathcal{E}^{\text{obs}}$. The following section proposes such a test, discusses the assumptions required to learn the set of d-invariant policies, and gives a detailed description of the whole procedure.

4. Learning an Optimal Invariant Policy

4.1. Learning invariant sets

Our theoretical results (Proposition 1 and Theorem 2) in the previous section assume that the set of all d-invariant policies Π^{inv} is given. We now turn to the task of learning Π^{inv} which boils down to searching for the collection of all d-invariant sets $\mathbf{S}^{\text{inv}}$ using data obtained from the observed environments $\mathcal{E}^{\text{obs}}$. To this end, we first define, for all $S \subseteq \{1, \dots, d\}$, $\pi \in \Pi$ and $\mathcal{E}' \subseteq \mathcal{E}$, the null hypothesis

$$H_0(S, \pi, \mathcal{E}') : \mathbb{P}^{\pi,e}_{R|X^S} \text{ is the same for all } e \in \mathcal{E}'. \tag{11}$$

In the case $\mathcal{E}' = \mathcal{E}^{obs}$, we refer to $H_0(S, \pi, \mathcal{E}^{obs})$ as $\mathcal{E}^{\text{obs}}$-invariance (which does not consider the unseen environments). Furthermore, we call a set S *invariant* if there exists $\pi \in \Pi^S$ such that $H_0(S, \pi, \mathcal{E}^{obs})$ holds and a policy π *invariant with respect to* S if $\pi \in \Pi^S$ and S is invariant. We now state our core assumptions that make learning possible.

Assumption 3. *For all $S \subseteq \{1, \dots, d\}$, the following holds:*

(i) $\exists \pi \in \Pi^S : H_0(S, \pi, \mathcal{E})$ *true* $\implies R \perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S$

(ii) $\forall \pi \in \Pi^S : H_0(S, \pi, \mathcal{E}^{\text{obs}})$ *true* $\implies H_0(S, \pi, \mathcal{E})$ *true*

Item 3(i) connects the conditional distribution invariance used in the null hypothesis (11) to the d-invariance condition given in (5) (The reversed implication follows by

Lemma D.1, Appendix D.3.3.) This assumption is a special case of the faithfulness assumption Pearl [2009] which is a fundamental assumption in causal discovery methods (e.g., Glymour et al. [2019]) that, in linear SCMs, holds with probability one if the linear coefficients are drawn from a distribution that is absolutely continuous with respect to Lebesgue measure Meek [1995], Spirtes et al. [2000]. Item 3(ii) ensures that any invariance found in the observed environments $\mathcal{E}^{\text{obs}}$ can be generalized to all environments $\mathcal{E}$. Implicitly, it requires that the observed environments are sufficiently heterogeneous[10]. This type of assumption is also at the core of other invariance-based methods Rojas-Carulla et al. [2018], Magliacane et al. [2018], Arjovsky et al. [2019], Pfister et al. [2021].

At first glance, Item 3(i) suggests that we have to check the hypothesis $H_0(S, \pi, \mathcal{E})$ for all $\pi \in \Pi^S$ to conclude whether or not S is d-invariant. Fortunately, as shown in Proposition 2, we actually only need to check the null hypothesis for a single $\pi \in \Pi^S$.

Proposition 2. *Assume Setting 1 and Assumption 3. Then, for all subsets $S \subseteq \{1, \dots, d\}$ and for all policies $\pi, \tilde{\pi} \in \Pi^S$, it holds that*

$$H_0(S, \pi, \mathcal{E}) \text{ true} \iff H_0(S, \tilde{\pi}, \mathcal{E}) \text{ true}. \tag{12}$$

Proof. See Appendix D.3.7. □

Assumption 3 and Proposition 2 make the learning problem tractable. The task of testing whether a set S is d-invariant boils down to testing the $\mathcal{E}^{\text{obs}}$-invariance hypothesis $H_0(S, \pi, \mathcal{E}^{\text{obs}})$ for a single $\pi \in \Pi^S$.

Testing $H_0(S, \pi, \mathcal{E}^{\text{obs}})$ for $\pi \in \Pi^S$ by directly checking for a change in the conditional distributions across environments in the observed data is, however, not in general possible. This is because the observed data may have been generated based on an initial policy π^0 that does not satisfy $\pi^0 \in \Pi^S$. It can therefore happen that $H_0(S, \pi, \mathcal{E}^{\text{obs}})$ is true but $H_0(S, \pi^0, \mathcal{E}^{\text{obs}})$ is not.

We illustrate this point using the example graph $\mathcal{G}$ given in Fig. 1b. For a policy depending only on $S = \{2\}$ the environment e is d-separated from R given $X^{\{2\}}$ in $\mathcal{G}^{\{2\}}$, which implies that $\{2\}$ is d-invariant, and in particular that $H_0(\{2\}, \pi^{\{2\}}, \mathcal{E}^{\text{obs}})$ is true by the Markov property (see Lemma D.1 in Appendix D.3.3). However, if the initial policy π^0 depends on both X^1 and X^2, then the path $e \to X^1 \to A \to R$ in Fig. 1b is open, which implies, by Assumption 3, that $H_0(\{2\}, \pi^{\{1,2\}}, \mathcal{E}^{\text{obs}})$ is not true. [11]

Thus, in general, we cannot directly test the $\mathcal{E}^{\text{obs}}$-invariance hypothesis of a set S by using the observed data that were generated by the initial policy. Instead, we need to test $H_0(S, \pi^S, \mathcal{E}^{\text{obs}})$ for a policy $\pi^S \in \Pi^S$ that is different from the data-generating policy π^0 (by Proposition 2 it suffices to test a single policy). As we detail in the following section, we can do so by applying an off-policy test for invariance by resampling the data to mimic the policy π^S.

[10] If the observed environments are identical, we clearly would not be able to find any d-invariant set and policy from the observed data. Item 3(ii) prevents such cases.

[11] In the same example, when conditioning on $\{1, 2\}$, the path $e \to X^1 \leftarrow U \to R$ is also open, which shows that $S = \{1, 2\}$ is not a d-invariant set.

4.2. Testing invariance under distributional shifts

Consider a set $S \subseteq \{1, \dots, d\}$ and a test policy $\pi^S \in \Pi^S$. To test the hypothesis $H_0(S, \pi^S, \mathcal{E}^{\text{obs}})$, we apply the off-policy test from Thams et al. [2021], which draws a target sample from π^S by resampling the offline data – drawn from π^0 – and then tests the invariance in this target sample. More formally, let $\mathcal{E}^{\text{obs}} := \{e_1, \dots, e_L\}$ and suppose that for every $e_j \in \mathcal{E}^{\text{obs}}$ a data set D^{e_j} consisting of n_e observations $D^{e_j} = \{(X_i^{e_j}, A_i^{e_j}, R_i^{e_j}, \pi^0(A_i^{e_j}|X_i^{e_j}))\}_{i=1}^{n_{e_j}}$ is available. For each environment e_j, we draw a weighted resample D^{e_j,π^S} of D^{e_j} using the weighted resampling procedure introduced in Thams et al. [2021].[12] We then apply an invariance test $\varphi^S(D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S})$ to the resampled data, to test the $\mathcal{E}^{\text{obs}}$-invariance hypothesis $H_0(S, \pi^S, \mathcal{E}^{\text{obs}})$. In Appendix D.5, we provide details on the resampling scheme, that is, a formal definition of D^{e_j,π^S} and show that the theoretical guarantees on the asymptotic level proved in Thams et al. [2021] also extend to our application. We detail a concrete test φ^S in Section 4.5 below.

4.3. Algorithm for invariant policy learning

The previous sections discuss finding invariant subsets S. We now discuss how to employ this in an algorithm that learns an optimal invariant policy. We assume that we are given an off-policy optimization algorithm `off\_opt` that takes as input a sample $D := (D^{e_1}, \dots, D^{e_L})$ and a policy space Π, and returns an optimal policy π^* and its estimated expected reward $\hat{\mathbb{E}}^{\pi^*}(R)$.

Here, we present one choice of `off\_opt` that we use in the experimental section; our approach can also be applied with other off-policy optimization algorithms. Given a policy space Π^S, we consider an optimal policy of the form

$$\pi^S(a \mid x) := \mathbb{1}\big[a = \arg\max_{a' \in \mathcal{A}} Q^S(x, a')\big], \tag{13}$$

where $Q^S(x, a) := \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}^{\pi_a, e_\ell}[R \mid X^S = x]$ denotes the pooled conditional mean under the policy that always selects an action a.

Let π^0 be an initial policy generating the sample D. By our assumption in Setting 1, the policy π^0 depends only on the observed covariates X. We, therefore, have that for all $S \subseteq \{1, \dots, d\}$ the pooled conditional mean $Q^S(x, a)$ is identifiable for all $a \in \mathcal{A}$ and $x \in \mathcal{X}^S$. We propose to estimate Q^S by a weighted least squares approach

$$\hat{Q}^S := \arg\min_{f \in \mathcal{F}^S} \sum_{\ell=1}^{L} \sum_{i=1}^{n_{e_\ell}} \frac{1}{\pi^0(A_i^{e_\ell} \mid X_i^{e_\ell})} (f(A_i^{e_\ell}, X_i^{e_\ell S}) - R_i^{e_\ell})^2,$$

where $\mathcal{F}^S \subseteq \{f : \boldsymbol{\mathcal{X}}^S \times \mathcal{A} \to \mathbb{R}\}$ is a class of functions. We then plug the estimate $\hat{Q}^S$ into (13) to obtain our (estimated) optimal policy.

If a subset S is found to be invariant, we can use `off\_opt` to learn an optimal policy

[12] Importance weighting is not applicable here because the test statistics of an invariance test cannot be expressed in terms of weighted averages. See also the discussion in Thams et al. [2021].

that uses S. Between all invariant subsets, we then select the one that has the highest estimated expected reward. We summarize the overall procedure for learning an optimal invariant policy, see Algorithm 1: The algorithm iterates over all subsets $S \subseteq \{1, \dots, d\}$ and checks the invariance condition using Algorithm 2 if one wants to use a known fixed test policy or Algorithm D.2 (see Appendix D.8) if one wants to use a test policy that optimizes the power of the invariance test, as described in Section 4.5.2. For each iteration, if the set S is invariant, we learn an optimal policy π^{S^*} within the policy space Π^S and compute its estimated expected reward $\hat{\mathbb{E}}^{\pi^{S^*}}(R)$ using `off\_opt`. Then, the algorithm returns an optimal policy π^{S^*} such that the estimated expected reward $\hat{\mathbb{E}}^{\pi^{S^*}}(R)$ is maximized. Lastly, the algorithm returns null if no invariant sets are found.

Algorithm 1 requires us to iterate over all subset $S \subseteq \{1, \dots, d\}$ which may be computationally intractable when d is large. We suggest two approaches for reducing the computational complexity of the algorithm. First, one can use a variable screening method (e.g., Lasso regression Tibshirani [1996]) to filter out the variables that are not predictive of the reward. If an optimal invariant set is a subset of the Markov blanket $\text{MB}(R)$ of the reward, applying a variable screening step prior to Algorithm 1 would not change the algorithm's output on the population level (see Peters et al. [2016], Rojas-Carulla et al. [2018], Pfister et al. [2021]). This approach is particularly efficient when the Markov blanket is sparse, that is, $|\text{MB}(R)| \ll d$.

Second, one may apply a greedy search instead of the exhaustive search in Algorithm 1. More specifically, we suggest to follow the greedy search introduced in Rojas-Carulla et al. [2018]. The greedy algorithm starts with an empty set $\hat{S} = \emptyset$. For each iteration, we search over the neighboring sets of the candidate set $\hat{S}$, which are obtained by adding or removing one predictor to or from $\hat{S}$. If any of the neighboring sets are accepted by the invaraince test, we select the one with the highest expected reward. If the test rejects all the neighbors, we select a neighbor that yields the largest p-value of the test.

4.4. Learning causal ancestors under distributional shifts

Sections 4.1 and 4.2 discuss an approach to learn invariant sets from off-policy data. The learned invariant sets are then used to find an optimal invariant policy as discussed in Section 4.3. Besides learning an optimal invariant policy, one can further use the proposed off-policy invariance test to analyze the causal structure. More specifically, the learned invariant sets allow us to look for potential observed causal ancestors $\text{AN}(R)$[13] of R by taking the intersection of the accepted sets. This approach is similar to invariant causal prediction Peters et al. [2016], except that here, we employ the off-policy invariance test to account for the distributional shift between the initial and the test policies, and allow for hidden variables.

Now we outline a method for finding $\text{AN}(R)$ from the offline data obtained from multiple environments $D^{e_1}, \dots, D^{e_L}$. For all $e_j \in \mathcal{E}^{\text{obs}}$ and $S \in \{1, \dots, d\}$, let us denote by D^{e_j, π^S} a weighted resample of D^{e_j}, and ψ^S an invariance test for the $\mathcal{E}^{\text{obs}}$-invariance

[13] Formally, $\text{AN}(R) \subseteq \{1, \dots, d\}$ is defined as the set of indicies j for which there is a directed path from X^j to R in $\mathcal{G}$.

Algorithm 1 Learning an optimal invariant policy

Input: data $D = (D^{e_1}, \dots, D^{e_L})$, off-policy optimization function `off\_opt`, test function `pv`, initial policy π_0, resampling size $\mathbf{m} := (m_1, \dots, m_L) = (\sqrt{|D^{e_1}|}, \dots, \sqrt{|D^{e_L}|})$

initialize maximum reward `maxR` $\leftarrow -\infty$
initialize optimal invariant policy $\pi^*_{\text{inv}} \leftarrow$ `null`
for $S \in \mathcal{P}(\{1, \dots, d\})$ **do** ▷ loop over all subsets
  if $\pi^S \neq$ `null` **then**
    `is_inv` $\leftarrow$ `test_inv`$(D, \pi^S, \texttt{pv}, S, \mathbf{m})$ ▷ see Algorithm 2
  else
    `is_inv` $\leftarrow$ `test_inv_opt_`$\pi(D, \texttt{pv}, S, \mathbf{m})$ ▷ (see Algorithm D.2 in Appendix D.8)
  if `is_inv` **then** ▷ update best invariant set
    $\pi^*_S, \hat{\mathbb{E}}^{\pi^*_S}(R) \leftarrow$ `off_opt`(D, Π^S)
    if `maxR` $< \hat{\mathbb{E}}^{\pi^*_S}(R)$ **then**
      `maxR` $\leftarrow \hat{\mathbb{E}}^{\pi^*_S}(R)$
      $\pi^*_{\text{inv}} \leftarrow \pi^*_S$

Output: optimal invariant policy π^*_{inv}

Algorithm 2 Testing the invariance of a set S with given test policy π^S

Function: `test_inv`(data $D = (D^{e_1}, \dots, D^{e_L})$, test policy π^S, function `pv` yielding the p-value of an invariance test, target set S, resampling size $(m_1, \dots, m_L) = (\sqrt{|D^{e_1}|}, \dots, \sqrt{|D^{e_L}|})$, significance level α ▷ resampling according to π^S

for $e = e_1, \dots, e_L$ **do**
  for $i = 1$ to $|D^e|$ **do**
    compute weights: $r^e_i \leftarrow \dfrac{\pi^S(a^e_i \mid x^{e,S}_i)}{\pi^0(a^e_i \mid x^e_i)}$
  draw $D^{e,\pi^S} := (D^e_{i_1}, \dots, D^e_{i_{m_e}})$ from D^e with prob. $\propto \prod_{\ell=1}^{m_e} r^e_{i_\ell}$
$D^{\pi^S} \leftarrow (D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S})$ ▷ verifying invariance condition
`is_invariant` $\leftarrow$ `pv`$(D^{\pi^S}) \geq \alpha$
Return `is_invariant`

hypothesis $H_0(S, \pi^S, \mathcal{E}^{\text{obs}})$ as discussed in Section 4.2 and Appendix D.5. For ease of presentation, we assume that $n_{e_1} = \dots = n_{e_L} =: n$. Then, we propose to estimate the causal ancestors of R by

$$\hat{S}^n_{\text{AN}} := \bigcap_{S : \psi^S(D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S}) = 0} S. \tag{14}$$

We detail the whole procedure in Algorithm D.1 in Appendix D.6. Proposition 3 shows

that this method controls the probability of wrongly selecting an incorrect variable.

Proposition 3. *Assume Setting 1, and that* $\mathbf{S}^{\text{inv}}$ *is non-empty. Let* $\hat{S}^n_{\text{AN}}$ *be the estimated set of causal ancestors given in* (14) *and assume that the invariance tests* ψ^S *used in* (14) *have pointwise asymptotic level* $\alpha \in (0, 1)$*. It then holds that*

$$\liminf_{n\to\infty} \mathbb{P}(\hat{S}^n_{\text{AN}} \subseteq \text{AN}(R)) \geq 1 - \alpha. \tag{15}$$

Proof. See Appendix D.3.8. □

4.5. Specifications of the target test

The resampling procedure detailed in Algorithm 2 requires a test function for the $\mathcal{E}^{\text{obs}}$-invariance null hypothesis that has power against the alternatives. We discuss one such test in Section 4.5.1 below. Moreover, in Sections 4.5.2 and 4.5.3, we discuss two choices of the test policy that aim to improve the power of the resampling test.

4.5.1. Invariant residual distribution test

We now detail a test φ^S to test $\mathcal{E}^{\text{obs}}$-invariance in the target sample. We first pool data from all environments into one data set and estimate the conditional $\mathbb{E}^{\pi^S}[R \mid X^S]$ using any prediction method (such as linear regression or a neural network). We then test whether the residuals $R - \mathbb{E}^{\pi^S}[R \mid X^S]$ are equally distributed across the environments $e \in \mathcal{E}$, i.e., we split the sample back into L groups (corresponding to the environments) and test whether the residuals in these groups are equally distributed (see also Peters et al. [2016], for example). We then define φ^S to be the composition of these operations, that is, φ^S returns 1 if the test for equal distribution of the residuals is rejected.

In the simulation and the warfarin case study (Section 5 and 6), we use the Kruskal-Wallis test Kruskal and Wallis [1952] to test whether the residuals have the same mean across environments; this test holds pointwise asymptotic level for all $\alpha \in (0, 1)$ (see Proposition D.2 in Appendix D.5). To obtain power against more alternatives, one could also use other tests, such as a two-sample kernel test with maximum mean discrepancy Gretton et al. [2012] and then correct for the multiple testing using Bonferroni-corrections (see also Rojas-Carulla et al. [2018], for example).

4.5.2. Optimizing the test policy for power

To check whether a subset S is invariant, we only need to test the $\mathcal{E}^{\text{obs}}$-invariance for a single policy $\pi \in \Pi^S$ (see Proposition 2). This provides us with a degree of freedom that we can leverage. Intuitively, the non-invariance may be more easily detectable in some test policies compared to others. We can therefore try to find a policy that gives us the strongest signal for detecting non-invariance. We maximize the power of the test by minimizing the p-value of the test. In a population setting, this would return small p-values for non-invariant sets, whereas for invariant sets one would not be able to make the p-values arbitrarily small, since they are uniformly distributed. In a finite sample

setting, this type of power optimization can lead to overfitting (which would break any level guarantees); to avoid this we use sample splitting.

As presented in Section 4.2, for each environment e, we obtain a target sample D^{e,π^S} from a test policy π^S by resampling the sample D^e that was generated under the policy π^0, and then test $\mathcal{E}^{\text{obs}}$-invariance in the target sample. The probabilities for obtaining the reweighted sample conditioned on the original sample are given by the importance weights, see Appendix D.5. Here, we optimize the ability to detect non-invariance over a parameterized subclass of Π^S,

$$\Pi_S^\Theta := \{\pi_\theta^S \mid \theta \in \Theta\},$$

where $\Theta = \bigtimes_{a\in\mathcal{A}} \mathbb{R}^{|S|}$ and π_θ^S is a linear softmax policy, i.e., for all $x^S \in \mathbb{R}^{|S|}$ and $a \in \mathcal{A}$:

$$\pi_\theta^S(a|x^S) = \frac{\exp\left(\theta_a^\top x^S\right)}{\sum_{a'} \exp\left(\theta_{a'}^\top x^S\right)}.$$

This is the parameterization we chose in the experiments below, but other choices work, too.

To check for the $\mathcal{E}^{\text{obs}}$-invariance condition of a subset S, the idea is then to find a policy $\pi_\theta^S \in \Pi_S^\Theta$ such that, in expectation, the test power is maximized, i.e., we need to solve the following optimization problem:

$$\arg\max_{\theta\in\Theta} \mathbb{E}\left[\texttt{pw}(D^{\pi_\theta^S}) \mid D\right],$$

where $D := (D^{e_1}, \ldots, D^{e_L})$ is all the observed data and $\texttt{pw}$ is a function that takes as input the reweighted sample $D^{\pi_\theta^S}$ and outputs the power of the test. Since we condition on D, the expectation is only with respect to the resampling of $D^{\pi_\theta^S}$. For many invariance tests, the test power $\texttt{pw}(D^{\pi_\theta^S})$ cannot be directly obtained, but one can minimize the p-value of the test instead. This motivates the objective function

$$\arg\min_{\theta\in\Theta} \mathbb{E}\left[\texttt{pv}(D^{\pi_\theta^S}) \mid D\right], \tag{16}$$

where $\texttt{pv}$ is a function that takes as input the reweighted sample $D^{\pi_\theta^S}$ and outputs the p-value of the test. We then employ gradient-based optimization algorithms to solve the above optimization problem, where the gradient is derived using the log-derivative. More precisely, let $J(\theta) := \mathbb{E}\left[\texttt{pv}(D^{\pi_\theta^S}) \mid D\right]$ be our objective function which now depends on

the parameters θ. The gradient of the objective function $J(\theta)$ can be derived as follows

$$\begin{aligned}\nabla J(\theta) &= \nabla \mathbb{E}\left[\mathtt{pv}(D^{\pi_\theta^S}) \mid D\right] \\ &= \nabla \sum_d \mathbb{P}(D^{\pi_\theta^S} = d \mid D)\mathtt{pv}(d) \\ &= \sum_d \mathbb{P}(D^{\pi_\theta^S} = d \mid D)\nabla \log \mathbb{P}(D^{\pi_\theta^S} = d \mid D)\mathtt{pv}(d) \\ &= \mathbb{E}\left[\nabla \log \mathbb{P}(D^{\pi_\theta^S} \mid D)\mathtt{pv}(D^{\pi_\theta^S}) \mid D\right].\end{aligned}$$

This expectation can be estimated by drawing repeated resamples $D^{\pi_\theta^S}$, where $\mathbb{P}(D^{\pi_\theta^S} \mid D)$ is determined by the resampling weights. In practice, we apply stochastic gradient descent Zhang [2004], i.e., at each iteration of the optimization we compute the gradient only from a single resample. As we argue in Appendix D.7, we can further speed up the optimization process substantially by a minor modification to the resampling weights, corresponding to sampling with replacement instead of distinct weights.

The optimization procedure yields a policy π_θ^* that approximately satisfies $\pi_\theta^* \in \arg\min_{\pi_\theta \in \Pi^S} J(\theta)$. We can then use π_θ^* as a test policy for testing the invariance of S. Lastly, to preserve the level of the statistical test, we split the original sample into two halves, perform the power optimization procedure on one half, and verify the invariance condition on the other half. The algorithm is presented in Algorithm D.2 in Appendix D.8. We only use the approximation of the resampling weights for the power optimization and use the actual resampling weights for the final resampling, so the level guarantee of Proposition D.2 in Appendix D.5 still holds.

4.5.3. Using a uniform target distribution

Since the procedure in Section 4.5.2 may be computationally challenging, especially if the algorithm is repeated many times as in Section 5. A computationally simpler approach is for each $a \in \mathcal{A}$ to test invariance under the test policy $\pi_a \in \Pi^{\emptyset}$, which always chooses the action a, and then combine the resulting p-values using Bonferroni corrections Dunn [1961]. Beyond computational simplicity, this has an additional benefit: Across environments there may be a cancelling effect of the difference in means due to different dependencies on the action in each environment. By testing the invariance of the conditional mean of the reward in each action, such cancelling effects are accounted for.

5. Simulation Experiments

To verify our theoretical findings we perform two simulation experiments, where we consider a linear multi-environment contextual bandit setting similar to Example 1 with

the following SCM $\mathcal{S}(\pi, e)$ (which induces the graph shown in Fig. 1b):

$$U \coloneqq \varepsilon_U, \quad X^1 \coloneqq \gamma_e U + \varepsilon_{X^1}, \quad X^2 \coloneqq \alpha_e + \varepsilon_{X^2},$$
$$A \sim \pi(A \mid X^1, X^2), \quad R \coloneqq \beta_{A,1} X^2 + \beta_{A,2} U + \varepsilon_R,$$

where $\varepsilon_U, \varepsilon_{X^1}, \varepsilon_{X^2}, \varepsilon_R \sim \mathcal{N}(0,1)$, A takes values in the space $\{a_1, \dots, a_L\}$, γ_e and α_e are parameters that depend on the environment e, and $\beta_{a_1,1}, \dots, \beta_{a_L,1}, \beta_{a_1,2}, \dots, \beta_{a_L,2}$ are parameters that are fixed across environments. Appendix D.9.1 contains details on how the parameters are chosen in the experiments. The code for all the experiments is available at https://github.com/sorawitj/invariant-policy-learning.

5.1. Generalization and invariance

We first consider an oracle setting, where we know a priori which subsets are invariant. From our data-generating process, it follows that $\{X^2\}$ is the only invariant set. We then compare an invariant policy which depends only on X^2 with a policy that uses both X^1 and X^2. We train both policies on a data set of size $10'000$ obtained from multiple training environments under a fixed initial policy π^0 (see Appendix D.9.2). In both cases, we employ a weighted least squares to estimate the expected reward $\mathbb{E}[R \mid A, X^S]$, where S is the subset that the policy uses. The policy then takes a greedy action w.r.t. the estimated expected reward, i.e., $\arg\max_a \hat{\mathbb{E}}[R \mid A = a, X^S]$ (see Section 4.3). Then we evaluate both policies on multiple unseen environments and compute the regret with respect to the policy that is optimal in each of the unseen environments. Fig. 2 shows the results. Each data point represents the evaluation on an unseen environment. The y-axes show the regret value and the x-axes display the distance from each unseen environment to the training environments (the distance is computed as the ℓ^2-distance between the average value of the pairs $(\gamma_{e_{tr}}, \alpha_{e_{tr}})$ in the training environments and the pair (γ_e, α_e) in the unseen test environment). The plot shows that the worst-case behavior of the invariant policy is smaller than the non-invariant one. In particular, for environments different from the training environments the gain can be significant. This empirically supports our result of Theorem 2.

5.2. Learning invariant policies

In practice, we do not know in advance which sets are invariant. We now aim to find an invariant policy from a data set generated under an initial policy π^0 which takes both X^1 and X^2 as input. To do so, we employ the method proposed in Section 4.2 for testing invariance under distributional shifts. More precisely, we generate a data set of size n from multiple training environments under the initial policy π^0 and apply the off-policy invariance test (see Section 4.5) to verify the invariance property of each subset in $\{\varnothing, \{X^1\}, \{X^2\}, \{X^1, X^2\}\}$. We repeat the experiment 500 times and plot the acceptance rates at various sample sizes ($n = 1'000, 3'000, 9'000, 27'000, 81'000$) (these numbers denote the total sample size, that is, number of observations, summed over all environments). The resulting acceptance rates are shown in Fig. 3. Our method yields

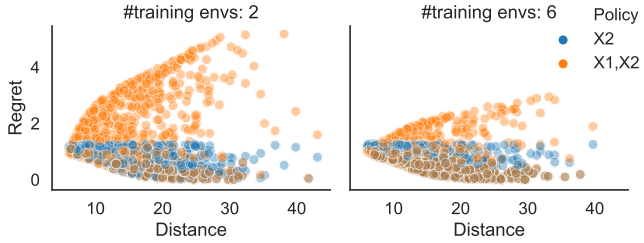


Figure 2.: The generalization performance (in terms of regret) of the policy based on an invariant set $\{X^2\}$ and the policy based on a non-invariant set $\{X^1, X^2\}$. The left and the right plot show the results when the training environments consist of two and six different environments, respectively. In both cases, the worst-case regret for the invariant policy is upper bounded while this is not the case for the non-invariant policy.

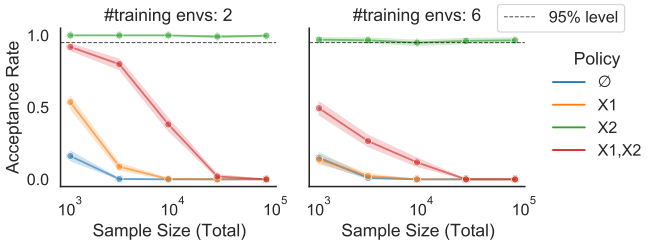


Figure 3.: Acceptance rates for the off-policy invariance test proposed in Section 4.2 for varying sample sizes. With increasing sample size, only the invariant set $\{X^2\}$ is accepted. Here, more environments (right) seems to yield higher test power than fewer environments (left).

high acceptance rates for the set $\{X^2\}$, which indeed is invariant, while the acceptance rates for other sets gradually decrease as the sample size increases. Furthermore, we can see that our test is more powerful when the number of training environments increases (keeping the total number of observations fixed). Our test is conservative (the acceptance rate is above the 95% level in the left plot) because the target test is not exact (the true conditional expectation is not given). In Appendix D.9.3, we conduct the same experiment with an exact test, using the true conditional expectation, which shows the correct level.

6. Warfarin Dosing Case Study

We evaluate our proposed approach on the clinical task of warfarin dosing. Warfarin is a blood thinner medicine prescribed to patients at risk of blood clots. The appropriate dose of warfarin varies from patient to patient depending on various factors such as

demographic and genetic information Consortium [2009]. Our case study is based on the International Warfarin Pharmacogenetics Consortium (IWPC) dataset Consortium [2009] which consists of 5′700 patients who were treated with warfarin, collected from 21 research groups on 4 continents. The IWPC dataset contains the optimal dose of warfarin for each of the patients as well as their information on demographic characteristics, clinical and genetic factors. The warfarin dosing problem has been used in a number of previous works evaluating off-policy learning algorithms Kallus and Zhou [2018], Bertsimas and McCord [2018], Zenati et al. [2020]. Similarly to these works, we formulate the warfarin dosing problem as a multi-environment contextual bandit problem as follows.

- The covariates (X) are patient-level features including demographic, clinical and genetic factors.
- The actions (A) are recommended warfarin doses output by a policy. We discretize the actions into three equal-sized buckets (low, medium, high) based on the quantiles of the optimal warfarin dose.
- The reward (R) depends on the recommended dose and the optimal dose: For each patient i, the reward $R_i(a)$ for an action $a \in \{\text{low}, \text{medium}, \text{high}\}$ is computed as

$$R_i(a) \coloneqq |Y_i - m(a)|\,, \tag{17}$$

  where Y_i is the optimal warfarin dose for a patient i and $m(a)$ is a median value of the optimal warfarin doses within the bucket a. Here, we assume that neither the reward function nor the optimal warfarin doses are known to the agent. Instead, for each patient i, only the reward for the action A_i is observed, i.e., $R_i \coloneqq R_i(A_i)$.
- The environments ($\mathcal{E}$) are proxies for continents. The continent information is not directly contained in the dataset, but we create proxies for the continent by clustering the 21 research groups into 4 clusters based on their proportion of the patients' race within each group. We believe that the resulting clusters roughly correspond to 4 different continents.

To reduce the search space, we select the top 10 features that are most predictive for the optimal warfarin dose using the permutation feature importance method Breiman [2001]. The top 10 features include 4 demographic variables, 4 clinical factors, and 2 genetic factors.

We consider two experimental setups to illustrate the benefits of our invariant learning approach. In the first setup, we directly apply our method to the IWPC dataset. Here, including invariance does not seem necessary in that our method performs similarly to other baselines (but not worse). It does, however, generate some causal insight into the problem. The second setup is a semi-real setting, where we introduce an artificial, non-invariant confounder.

We now outline our first experimental setup and the results. We first generate training data $\{(X_i, A_i, R_i, e_i)\}_{i=1}^n$ by drawing actions A_i from a policy $\pi^0 \in \Pi^{\text{BMI}}$ that is constructed from linear regression $Y_i \approx f(X_i^{\text{BMI}})$ of the optimal dose onto the BMI (see Appendix D.10.1 for more details).

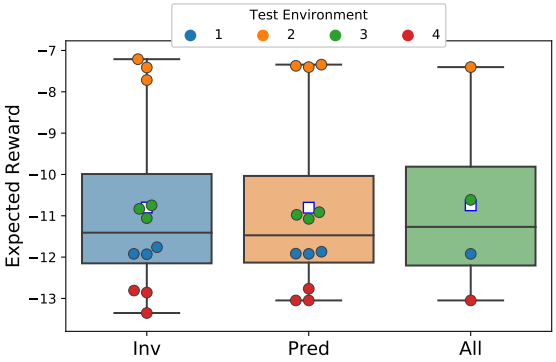


Figure 4.: Empirical results on the original data set. Each point represents the expected reward of a policy on the corresponding test environment. The square points represent the mean value of the expected rewards. In this setup, all candidate methods yield similar performances on all of the test environments. This result indicates that here predictive models are also invariant.

6.1. Candidate methods

Using the generated training data, we empirically compare the performance of the following policy learning methods:

- Invariant Policy Learning (Inv): This is our proposed method. We first perform the off-policy invariance test using the test described in Section 4.5.3 to search for potential invariant sets. We then take the top 20 sets with the largest p-values $\mathbf{S}_{\text{inv}}^{20}$ as the candidate invariant sets. For each S in $\mathbf{S}_{\text{inv}}^{20}$, we fit the policy optimization algorithm described in Section 4.3 with X^S as the covariates (the same algorithm is also used in other candidate methods below). Lastly, we select the top 3 sets that yield the largest expected rewards (computed using 5-fold cross-validation).
- Predictive Policy Learning (Pred): This method serves as a baseline for policy learning that solely maximizes the expected reward. For each subset S, we fit the policy optimization algorithm with X^S as the covariates. We then take the policies corresponding to the top 3 sets with the largest expected rewards.
- All Set Policy Learning (All): This method serves as another baseline where we take all of the patient's features and fit the policy optimization algorithm.

6.2. Evaluation setup & results

We compare the policy learning methods using the following 'leave-one-environment-out' evaluation procedure.

1. Select $e \in \mathcal{E} = \{1, \ldots, 4\}$ as a test environment. Split the training data into $D^{\text{test}} := \{(X_i, A_i, R_i, e_i)\}_{i=1}^{n_{test}}$, where $e_i = e$ and $D^{\text{tr}} := \{(X_i, A_i, R_i, e_i)\}_{i=1}^{n_{tr}}$, where $e_i \in \{1, \ldots, 4\} \setminus \{e\}$.

2. Using D^{tr}, train the policies with candidate methods detailed in Section 6.1.

3. Evaluate the fitted policies by computing the expected reward on D^{test} using the true reward function (17).

We repeat the above procedure for each $e \in \mathcal{E}$ and display the evaluation result in Fig. 4. The performances of all candidate methods are similar. Even though the proposed invariant approach does not yield a higher reward compared with the baselines, it does not worsen the performance, either. This suggests that we can gain the stability benefit of an invariant policy without having to sacrifice predictiveness. Indeed, the stability benefit could prevent the learned invariant policy from being suboptimal when a new test environment is sufficiently different from the training environments as we show in Section 6.4

6.3. Analyzing invariant sets

In addition to learning an optimal invariant policy, we can use the invariance-based approach to further analyze the dependence between the patient's features and the reward as discussed in Section 4.4. In particular, we apply the off-policy invariant causal prediction algorithm (see Algorithm D.1 in Appendix D.6) to find potential causal ancestors of the reward. On this dataset, with a confidence level of 5%, the algorithm returns the empty set, which can happen if the covariates are highly correlated, for example Heinze-Deml et al. [2018]. Nonetheless, we can still extract more information by obtaining the defining sets (see Section 2.2 in Heinze-Deml et al. [2018]). The resulting defining set of size 2 is {Race, VKORC1} (see Appendix D.10.2 for more details on the variables). These variables are potential causal ancestors in the sense that at least one variable in these sets is a causal ancestor.

6.4. Semi-real experiment

To further illustrate the benefits of the invariance-based learning approach, we consider a semi-real setup where we introduce hidden variables and a non-invariant predictor. We remove the two genetic factors from the patient's features and create a non-invariant predictor that depends on those two factors as follows.

We first fit a linear regression to estimate the optimal warfarin dose from the genetic factors and denote the resulting coefficients by β. To mimic environmental perturbations, we perturb β depending an environment $e \in \mathcal{E}$ resulting in $\beta_e := \gamma_e \beta$, where γ_e is an environment-specific parameter. We define the non-invariant predictor in the environment $e \in \mathcal{E}$ as $X^{\text{n-inv}} := X^{G^\top} \beta_e$, where X^G are the two genetic features. We then add $X^{\text{n-inv}}$ as part of the patient's features and remove X^G. The training data are generated in a similar fashion as in the first setup, except that the initial policy does not only depend on the BMI score X^{BMI} but also on the non-invariant predictor $X^{\text{n-inv}}$.

In addition to the candidate methods described in Section 6.1, we introduce an additional baseline for this setup.

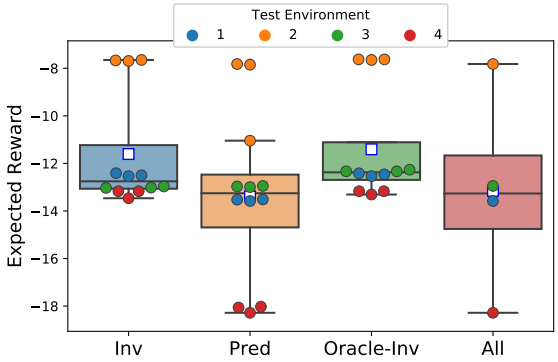


Figure 5.: Empirical results on policy learning with a non-invariant predictor (see Section 6.4). Each point represents the expected reward of a policy on the corresponding test environment. In this setup, our proposed method (Inv) outperforms the two baselines (Pred and All) that ignore the environment structure, while approaching the performance of the invariant oracle (Oracle-Inv).

- Oracle invariant Policy (Oracle-Inv): By construction, we know that $X^{\text{n-inv}}$ is a strongly non-d-invariant variable (see Definition 5). This method serves as an oracle version of the invariant policy learning method by searching for the top 3 sets that do not contain $X^{\text{n-inv}}$ such that their corresponding policies yield the largest expected reward (the procedure is similar to the Pred method with $X^{\text{n-inv}}$ being removed).

We evaluate the candidate methods using a similar procedure as described in Section 6.2. Fig. 5 illustrates the evaluation result. Our proposed method (Inv) yields a higher expected reward than the two baselines on most of the test environments. This is because the two baselines ignore the environment structure and use information from $X^{\text{n-inv}}$ in their resulting policies, while the invariant method uses the invariance test to remove this non-invariant proxy variable. Furthermore, the performance of our proposed method is almost on par with the invariant oracle (Oracle-Inv), except for the test environment $e = 3$, in which our approach is unable to ignore the non-invariant predictor, possibly because the non-invariance that would be implied by Assumption 3 may not be strong enough (for our test) when $\mathcal{E}^{\text{obs}} = \{1, 2, 4\}$.

7. Conclusion

This paper tackles the problem of environmental shifts in offline contextual bandits from a causal perspective. We introduce a framework for multi-environment contextual bandits that is based on structural causal models and frame the environmental shift problem as a distributionally robust objective over environments that are induced by different perturbations on the covariates. We prove that if there are no unobserved confounders, taking into account causality and invariance is not necessary for obtaining the distribu-

tionally robust policies. However, causality and invariance can become relevant when not all variables are observed. To tackle settings with unobserved confounders, we adapt invariance-based ideas from causal inference to the proposed framework and introduce the notion of invariant policies. Our theoretical results show that under certain assumptions an invariant policy that is optimal on the training environments is also optimal on all unseen environments, and therefore distributionally robust. We further provide a method for finding invariant policies based on an off-policy invariance test. It can be combined with any existing policy optimization algorithm to learn an optimal invariant policy. We believe that our contributions shed some light on what causality can offer in contextual bandit and, more generally, in reinforcement learning problems.

For future work, there are several directions that would be interesting to investigate. One direction is to explore the use of invariance-based ideas in the adaptive setting, in which the goal of an agent is to optimally adapt to a changing environment. Learning agents may require fewer and safer explorations in a new environment if they carry over invariance information from previous environments. It may further be possible to extend invariance-based ideas from the contextual bandit setting to the full reinforcement learning problem with long-term consequences and state dynamics. Although some previous works have explored this direction Zhang et al. [2020], Sonar et al. [2021], we believe that the connections with respect to causality and invariance are not yet fully understood. In the i.i.d. setting, recent work has investigated trading off invariance and predictability Rothenhäusler et al. [2021], Pfister et al. [2021], Jakobsen and Peters [2021], Oberst et al. [2021], Saengkyongam et al. [2022]. We believe that a similar idea can be applied to contextual bandit and reinforcement learning problems. Lastly, if one can gain additional knowledge of the test environments, one may aim to optimize objectives other than the worst-case performance which could lead to a different class of generalization guarantees.

This paper considers invariance as a dichotomous property and could be a first step towards using invariance-based ideas for building safer and more robust adaptive learning systems.

Acknowledgments

SS, NT, and JP were supported by a research grant (18968) from VILLUM FONDEN and JP was, in addition, supported by the Carlsberg Foundation. NP was supported by a research grant (0069071) from Novo Nordisk Fonden. We thank Steffen Lauritzen for helpful discussions.

4. Shifts in Distributions: Causal Inference

This chapter contains the following two papers:

[TimeIV] [Thams et al., 2022b]. N. Thams, R. Søndergaard, S. Weichwald, and J. Peters. Identifying causal effects using instrumental time series: Nuisance IV and correcting for the past. *arXiv preprint arXiv:2203.06056*, 2022b.

[AncSearch] [Mogensen et al., 2022]. P. Mogensen, N. Thams, and J. Peters. Invariant ancestry search. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15832–15857. PMLR, 2022.

In this chapter we develop methods for drawing causal inference, that is, for inferring properties of the structural causal model which generated the data. Both papers draw their identifiability results from the presence of exogenous information, either from different environments or from an instrumental variable.

In **[AncSearch]**, we assume that we observe data in several environments, each of which corresponds to a change in distribution of one or more of the observed covariates, similar to the setup in ICP (see Section 1.3 in Chapter 1). A drawback of ICP, which for a target variable Y uses invariance of conditionals to learn a subset of causal parents, is that in many cases the learned subset is small or even empty. This occurs because ICP outputs the intersection over all invariant sets of predictors, so if disjoint invariant sets exist, the learned subset is empty. The solution we propose learns a larger set of causal ancestors than ICP, by defining 'minimal invariance', and taking the union over all minimally invariant sets of predictors. This comes at the cost that the learned set not only contains parents but also other ancestors; additionally while ICP has finite-sample level guarantees, our approach also relies on having power to reject a hypothesis of invariance and therefore only has asymptotic guarantees.

In **[TimeIV]**, we consider IV models in a linear time series, where we observe an exogenous instrument processes, and want to estimate the causal effect X_{t-1} on Y_t, where $(X_s, Y_s)_{s\in\mathbb{Z}}$ are time series that are confounded by an unobserved time series. A difficulty is that the instruments and the response variable correlate not only because of direct dependencies, but also because of past states of the time series, which can cause the IV assumptions (see Section 1.2 in Chapter 1) to fail. We show that by using conditional IV [Pearl, 2009] we can adjust for this 'confounding from the past'. Even with a valid instrument, identification of the causal effect requires that the instruments are sufficiently high-dimensional. We show that even if this instrument process is low-dimensional (which in IV for i.i.d. data means that the causal effect may not be

identifiable) we can use past states of the time series as additional instruments, and thereby identify causal effects.

Invariant Ancestry Search

Phillip B. Mogensen, Nikolaj Thams and Jonas Peters

Abstract

Recently, methods have been proposed that exploit the invariance of prediction models with respect to changing environments to infer subsets of the causal parents of a response variable. If the environments influence only few of the underlying mechanisms, the subset identified by invariant causal prediction (ICP), for example, may be small, or even empty. We introduce the concept of minimal invariance and propose invariant ancestry search (IAS). In its population version, IAS outputs a set which contains only ancestors of the response and is a superset of the output of ICP. When applied to data, corresponding guarantees hold asymptotically if the underlying test for invariance has asymptotic level and power. We develop scalable algorithms and perform experiments on simulated and real data.

1. Introduction

Causal reasoning addresses the challenge of understanding why systems behave the way they do and what happens if we actively intervene. Such mechanistic understanding is inherent to human cognition, and developing statistical methodology that learns and utilizes causal relations is a key step in improving both narrow and broad AI [Jordan, 2019, Pearl, 2018]. Several approaches exist for learning causal structures from observational data. Approaches such as the PC-algorithm [Spirtes et al., 2000] or greedy equivalence search [Chickering, 2002] learn (Markov equivalent) graphical representations of the causal structure Lauritzen [1996]. Other approaches learn the graphical structure under additional assumptions, such as non-Gaussianity Shimizu et al. [2006] or non-linearity Hoyer et al. [2009], Peters et al. [2014]. Zheng et al. [2018] convert the problem into a continuous optimization problem, at the expense of identifiability guarantees.

Invariant causal prediction (ICP) [Peters et al., 2016, Heinze-Deml et al., 2018, Pfister et al., 2019b, Gamella and Heinze-Deml, 2020, Martinet et al., 2022] assumes that data are sampled from heterogeneous environments (which can be discrete, categorical or continuous), and identifies direct causes of a target Y, also known as causal parents of Y. Learning ancestors (or parents) of a response Y yields understanding of anticipated changes when intervening in the system. It is a less ambitious task than learning the

complete graph but may allow for methods that come with weaker assumptions and stronger guarantees. More concretely, for predictors $X_1, \dots, X_d$, ICP searches for subsets $S \subseteq \{1, \dots, d\}$ that are invariant; a set X_S of predictors is called invariant if it renders Y independent of the environment, conditional on X_S. ICP then outputs the intersection of all invariant predictor sets $S_{\text{ICP}} \coloneqq \cap_{S \text{ invariant}} S$. Peters et al. [2016] show that if invariance is tested empirically from data at level α, the resulting intersection $\hat{S}_{\text{ICP}}$ is a subset of direct causes of Y with probability at least $1 - \alpha$.[1]

In many cases, however, the set learned by ICP forms a strict subset of all direct causes or may even be empty. This is because disjoint sets of predictors can be invariant, yielding an empty intersection, which may happen both for finite samples as well as in the population setting. In this work, we introduce and characterize minimally invariant sets of predictors, that is, invariant sets S for which no proper subset is invariant. We propose to consider the union S_{IAS} of all minimally invariant sets, where IAS stands for invariant ancestry search. We prove that S_{IAS} is a subset of causal ancestors of Y, invariant, non-empty and contains S_{ICP}. Learning causal ancestors of a response may be desirable for several reasons: e.g., they are the variables that may have an influence on the response variable when intervened on. In addition, because IAS yields an invariant set, it can be used to construct predictions that are stable across environments [e.g., Rojas-Carulla et al., 2018, Christiansen et al., 2021].

In practice, we estimate minimally invariant sets using a test for invariance. If such a test has asymptotic power against some of the non-invariant sets (specified in Section 5.2), we show that, asymptotically, the probability of $\hat{S}_{\text{IAS}}$ being a subset of the ancestors is at least $1 - \alpha$. This puts stronger assumptions on the invariance test than ICP (which does not require any power) in return for discovering a larger set of causal ancestors. We prove that our approach retains the ancestral guarantee if we test minimal invariance only among subsets up to a certain size. This yields a computational speed-up compared to testing minimal invariance in all subsets, but comes at the cost of potentially finding fewer causal ancestors.

The remainder of this work is organized as follows. In Section 2 we review relevant background material, and we introduce the concept of minimal invariance in Section 3. Section 4 contains an oracle algorithm for finding minimally invariant sets (and a closed-form expression of S_{ICP}) and Section 5 presents theoretical guarantees when testing minimal invariance from data. In Section 6 we evaluate our method in several simulation studies as well as a real-world data set on gene perturbations. Code is provided at `https://github.com/PhillipMogensen/InvariantAncestrySearch`.

[1] Rojas-Carulla et al. [2018], Magliacane et al. [2018], Arjovsky et al. [2019], Christiansen et al. [2021] propose techniques that consider similar invariance statements with a focus on distribution generalization instead of causal discovery.

2. Preliminaries

2.1. Structural causal models and graphs

We consider a setting where data are sampled from a structural causal model (SCM) Pearl [2009], Bongers et al. [2021]

$$Z_j := f_j(\mathrm{PA}_j, \varepsilon_j),$$

for some functions f_j, parent sets PA_j and noise distributions ε_j. Following Peters et al. [2016], Heinze-Deml et al. [2018], we consider an SCM over variables $Z := (E, X, Y)$ where E is an exogenous environment variable (i.e., $\mathrm{PA}_E = \emptyset$), Y is a response variable and $X = (X_1, \ldots, X_d)$ is a collection of predictors of Y. We denote by $\mathcal{P}$ the family of all possible distributions induced by an SCM over (E, X, Y) of the above form.

For a collection of nodes $j \in [d] := \{1, \ldots, d\}$ and their parent sets PA_j, we define a directed graph $\mathcal{G}$ with nodes $[d]$ and edges $j' \to j$ for all $j' \in \mathrm{PA}_j$. We denote by CH_j, AN_j and DE_j the children, ancestors and descendants of a variable j, respectively, neither containing j. A graph $\mathcal{G}$ is called a directed acyclic graph (DAG) if it does not contain any directed cycles. See Pearl [2009] for more details and the definition of d-separation.

Throughout the remainder of this work, we make the following assumptions about causal sufficiency and exogeneity of E (Section 7 describes how these assumptions can be relaxed).

Assumption 1. *Data are sampled from an SCM over nodes (E, X, Y), such that the corresponding graph is a DAG, the distribution is faithful with respect to this DAG, and the environments are exogenous, i.e.,* $\mathrm{PA}_E = \emptyset$.

2.2. Invariant causal prediction

Invariant causal prediction (ICP), introduced by Peters et al. [2016], exploits the existence of heterogeneity in the data, here encoded by an environment variable E, to learn a subset of causal parents of a response variable Y. A subset of predictors $S \subseteq [d]$ is *invariant* if $Y \perp\!\!\!\perp E \mid S$, and we define $\mathcal{I} := \{S \subseteq [d] \mid S \text{ invariant}\}$ to be the set of all invariant sets. We denote the corresponding hypothesis that S is invariant by

$$H_{0,S}^{\mathcal{I}}: \quad S \in \mathcal{I}.$$

Formally, $H_{0,S}^{\mathcal{I}}$ corresponds to a subset of distributions in $\mathcal{P}$, and we denote by $H_{A,S}^{\mathcal{I}} := \mathcal{P} \setminus H_{0,S}^{\mathcal{I}}$ the alternative hypothesis to $H_{0,S}^{\mathcal{I}}$. Peters et al. [2016] define the oracle output

$$S_{\mathrm{ICP}} := \bigcap_{S: H_{0,S}^{\mathcal{I}} \text{ true}} S \tag{1}$$

(with $S_{\mathrm{ICP}} = \emptyset$ if no sets are invariant) and prove $S_{\mathrm{ICP}} \subseteq \mathrm{PA}_Y$. If provided with a test for the hypotheses $H_{0,S}^{\mathcal{I}}$, we can test all sets $S \subseteq [d]$ for invariance and take the

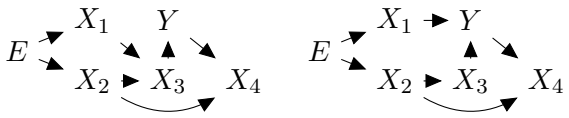


Figure 1.: Two structures where $S_{\text{ICP}} \subsetneq \text{PA}_Y$. (*left*) $S_{\text{ICP}} = \emptyset$. (*right*) $S_{\text{ICP}} = \{1\}$. In both, our method outputs $S_{\text{IAS}} = \{1, 2, 3\}$.

intersection over all accepted sets: $\hat{S}_{\text{ICP}} := \bigcap_{S: H_{0,S}^{\mathcal{I}} \text{ not rejected}} S$; If the invariance test has level α, $\hat{S}_{\text{ICP}} \subseteq \text{PA}_Y$ with probability at least $1 - \alpha$.

However, even for the oracle output in (1), there are many graphs for which S_{ICP} is a strict subset of PA_Y. For example, in Fig. 1 (left), since both $\{1, 2\}$ and $\{3\}$ are invariant, $S_{\text{ICP}} \subseteq \{1, 2\} \cap \{3\} = \emptyset$. This does not violate $S_{\text{ICP}} \subseteq \text{PA}_Y$, but is non-informative. Similarly, in Fig. 1 (right), $S_{\text{ICP}} = \{1\}$, as all invariant sets contain $\{1\}$. Here, S_{ICP} contains some information, but is not able to recover the full parental set. In neither of these two cases, S_{ICP} is an invariant set. If the environments are such that each parent of Y is either affected by the environment directly or is a parent of an affected node, then $S_{\text{ICP}} = \text{PA}_Y$ [Peters et al., 2016, proof of Theorem 3]. The shortcomings of ICP thus relate to settings where the environments act on too few variables or on uninformative ones.

For large d, it has been suggested to apply ICP to the variables in the *Markov boundary* Pearl [2014], $\text{MB}_Y = \text{PA}_Y \cup \text{CH}_Y \cup \text{PA}(\text{CH}_Y)$ (we denote the oracle output by $S_{\text{ICP}}^{\text{MB}}$). As $\text{PA}_Y \subseteq \text{MB}_Y$, it still holds that $S_{\text{ICP}}^{\text{MB}}$ is a subset of the causal parents of the response.[2] However, the procedure must still be applied to $2^{|\,\text{MB}_Y\,|}$ sets, which is only feasible if the Markov boundary is sufficiently small. In practice, the Markov boundary can, for example, be estimated using Lasso regression or gradient boosting techniques [Tibshirani, 1996, Meinshausen and Bühlmann, 2006, Friedman, 2001].

3. Minimal Invariance and Ancestry

We now introduce the concept of minimally invariant sets, which are invariant sets that do not have any invariant subsets. We propose to consider S_{IAS}, the oracle outcome of invariant ancestry search, defined as the union of all minimally invariant sets. We will see that S_{IAS} is an invariant set, it consists only of ancestors of Y, and it contains S_{ICP} as a subset.

Definition 1. Let $S \subseteq [d]$. We say that S is *minimally invariant* if and only if

$$S \in \mathcal{I} \text{ and } \forall S' \subsetneq S: \; S' \notin \mathcal{I};$$

that is, S is invariant and no subset of S is invariant. We define $\mathcal{MI} := \{S \mid S \text{ minimally invariant}\}$.

[2] In fact, $S_{\text{ICP}}^{\text{MB}}$ is always at least as informative as ICP. E.g., there exist graphs in which $S_{\text{ICP}} = \emptyset$ and $S_{\text{ICP}}^{\text{MB}} \neq \emptyset$, see Fig. 1 (left). There are no possible structures for which $S_{\text{ICP}}^{\text{MB}} \subsetneq S_{\text{ICP}}$, as both search for invariant sets over all sets of parents of Y.

The concept of minimal invariance is closely related to the concept of minimal d-separators [Tian et al., 1998]. This connection allows us to state several properties of minimal invariance. For example, an invariant set is minimally invariant if and only if it is non-invariant as soon as one of its elements is removed.

Proposition 1. *Let $S \subseteq [d]$. Then $S \in \mathcal{MI}$ if and only if $S \in \mathcal{I}$ and for all $j \in S$, it holds that $S \setminus \{j\} \notin \mathcal{I}$.*

The proof follows directly from [Tian et al., 1998, Corollary 2]. We can therefore decide whether a given invariant set S is minimally invariant using $\mathcal{O}(|S|)$ checks for invariance, rather than $\mathcal{O}(2^{|S|})$ (as suggested by Definition 1). We use this insight in Section 5.1, when we construct a statistical test for whether or not a set is minimally invariant.

To formally define the oracle outcome of IAS, we denote the hypothesis that a set S is minimally invariant by

$$H_{0,S}^{\mathcal{MI}}: \quad S \in \mathcal{MI}$$

(and the alternative hypothesis, $S \notin \mathcal{MI}$, by $H_{A,S}^{\mathcal{MI}}$) and define the quantity of interest

$$S_{\text{IAS}} := \bigcup_{S: H_{0,S}^{\mathcal{MI}} \text{ true}} S \tag{2}$$

with the convention that a union over the empty set is the empty set.

The following proposition states that S_{IAS} is a subset of the ancestors of the response Y. Similarly to PA_Y, variables in AN_Y are causes of Y in that for each ancestor there is a directed causal path to Y. Thus, generically, when intervened, these variables have a causal effect on the response.

Proposition 2. *It holds that $S_{\text{IAS}} \subseteq \text{AN}_Y$.*

The proof follows directly from [Tian et al., 1998, Theorem 2]; see also [Acid and De Campos, 2013, Proposition 2]. The setup in these papers is more general than what we consider here; we therefore provide direct proofs for Propositions 1 and 2 in Appendix E.1, which may provide further intuition for the results.

Finally, we show that the oracle output of IAS contains that of ICP and, contrary to ICP, it is always an invariant set.

Proposition 3. *Assume that $E \notin \text{PA}_Y$. It holds that*

(i) $S_{\text{IAS}} \in \mathcal{I}$ and

(ii) $S_{\text{ICP}} \subseteq S_{\text{IAS}}$, with equality if and only if $S_{\text{ICP}} \in \mathcal{I}$.

4. Oracle Algorithms

When provided with an oracle that tells us whether a set is invariant or not, how can we efficiently compute S_{ICP} and S_{IAS}? Here, we assume that the oracle is given by a

DAG, see Assumption 1. A direct application of (1) and (2) would require checking a number of sets that grows exponentially in the number of nodes. For S_{ICP}, we have the following characterization.[3]

Proposition 4. *If* $E \notin \text{PA}_Y$, *then* $S_{\text{ICP}} = \text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E))$.

This allows us to efficiently read off S_{ICP} from the DAG, (e.g., it can naively be done in $\mathcal{O}((d+2)^{2.373} \log(d+2))$ time, where the exponent 2.373 comes from matrix multiplication). For S_{IAS}, to the best of our knowledge, there is no closed form expression that has a similarly simple structure.

Instead, for IAS, we exploit the recent development of efficient algorithms for computing all minimal d-separators (for two given sets of nodes) in a given DAG [see, e.g., Tian et al., 1998, van der Zander et al., 2019]. A set S is called a *minimal d-separator* of E and Y if it d-separates E and Y given S and no strict subset of S satisfies this property. These algorithms are often motivated by determining minimal adjustment sets [e.g., Pearl, 2009] that can be used to compute the total causal effect between two nodes, for example. If the underlying distribution is Markov and faithful with respect to the DAG, then a set S is minimally invariant if and only if it is a minimal d-separator for E and Y. We can therefore use the same algorithms to find minimally invariant sets; van der Zander et al. [2019] provide an algorithm (based on work by Takata [2010]) for finding minimal d-separators with polynomial delay time. Applied to our case, this means that while there may be exponentially many minimally invariant sets,[4] when listing all such sets it takes at most polynomial time until the next set or the message that there are nor further sets is output. In practice, on random graphs, we found this to work well (see Section 6.1). But since S_{IAS} is the union of all minimally invariant sets, even faster algorithms may be available; to the best of our knowledge, it is an open question whether finding S_{IAS} is an NP-hard problem (see Appendix E.2 for details).

We provide a function for listing all minimally invariant sets in our python code; it uses an implementation of the above mentioned algorithm, provided in the R [R Core Team, 2021] package `dagitty` [Textor et al., 2016]. In Section 6.1, we study the properties of the oracle set S_{IAS}. When applied to 500 randomly sampled, dense graphs with $d = 15$ predictor nodes and five interventions, the `dagitty` implementation had a median speedup of a factor of roughly 17, compared to a brute-force search (over the ancestors of Y). The highest speedup achieved was by a factor of more than 1,900.

The above mentioned literature can be used only for oracle algorithms, where the graph is given. In the following sections, we discuss how to test the hypothesis of minimal invariance from data.

[3] To the best of our knowledge, this characterization is novel.

[4] This is the case if there are $d/2$ (disjoint) directed paths between E and Y, with each path containing two X-nodes, for example [e.g., van der Zander et al., 2019].

5. Invariant Ancestry Search

5.1. Testing a single set for minimal invariance

Usually, we neither observe a full SCM nor its graphical structure. Instead, we observe data from an SCM, which we want to use to decide whether a set is in $\mathcal{MI}$, such that we make the correct decision with high probability. We now show that a set S can be tested for minimal invariance with asymptotic level and power if given a test for invariance that has asymptotic level and power.

Assume that $\mathcal{D}_n = (X_i, E_i, Y_i)_{i=1}^n$ are observations (which may or may not be independent) of (X, E, Y) and let $\phi_n^{\mathcal{MI}} : \text{powerset}([d]) \times \mathcal{D}_n \times (0,1) \to \{0,1\}$ be a decision rule that transforms $(S, \mathcal{D}_n, \alpha)$ into a decision $\phi_n^{\mathcal{MI}}(S, \mathcal{D}_n, \alpha)$ about whether the hypothesis $H_{0,S}^{\mathcal{MI}}$ should be rejected ($\phi_n^{\mathcal{MI}} = 1$) at significance threshold α, or not ($\phi_n^{\mathcal{MI}} = 0$). To ease notation, we suppress the dependence on $\mathcal{D}_n$ and α when the statements are unambiguous.

A test ψ_n for the hypothesis H_0 has pointwise asymptotic level if

$$\forall \alpha \in (0,1): \quad \sup_{\mathbb{P} \in H_0} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) \leq \alpha \tag{3}$$

and pointwise asymptotic power if

$$\forall \alpha \in (0,1): \quad \inf_{\mathbb{P} \in H_A} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) = 1. \tag{4}$$

If the limit and the supremum (resp. infimum) in (3) (resp. (4)) can be interchanged, we say that ψ_n has uniform asymptotic level (resp. power).

Tests for invariance have been examined in the literature. Peters et al. [2016] propose two simple methods for testing for invariance in linear Gaussian SCMs when the environments are discrete, although the methods proposed extend directly to other regression scenarios. Pfister et al. [2019b] propose resampling-based tests for sequential data from linear Gaussian SCMs. Furthermore, any valid test for conditional independence between Y and E given a set of predictors S can be used to test for invariance. Although for continuous X, there exists no general conditional independence test that has both level and non-trivial power [Shah and Peters, 2020], it is possible to impose restrictions on the data-generating process that ensure the existence of non-trivial tests [e.g., Fukumizu et al., 2008, Zhang et al., 2011, Berrett et al., 2020, Shah and Peters, 2020, Thams et al., 2021]. Heinze-Deml et al. [2018] provide an overview and a comparison of several conditional independence tests in the context of invariance.

To test whether a set $S \subseteq [d]$ is minimally invariant, we define the decision rule

$$\phi_n^{\mathcal{MI}}(S) := \begin{cases} 1 & \text{if } \phi_n(S) = 1 \text{ or } \min_{j \in S} \phi_n(S \setminus \{j\}) = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\phi_n^{\mathcal{MI}}(\emptyset) := \phi_n(\emptyset)$. Here, ϕ_n is a test for the hypothesis $H_{0,S}^{\mathcal{I}}$, e.g., one of the tests

mentioned above. This decision rule rejects $H_{0,S}^{\mathcal{MI}}$ either if $H_{0,S}^{\mathcal{I}}$ is rejected by ϕ_n or if there exists $j \in S$ such that $H_{0,S\setminus\{j\}}^{\mathcal{I}}$ is not rejected. If ϕ_n has pointwise (resp. uniform) asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has pointwise (resp. uniform) asymptotic level and pointwise (resp. uniform) asymptotic power of at least $1-\alpha$.

Theorem 1. *Let $\phi_n^{\mathcal{MI}}$ be defined as in (5) and let $S \subseteq [d]$. Assume that the decision rule ϕ_n has pointwise asymptotic level and power for S and for all $S\setminus\{j\}, j \in S$. Then, $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level and pointwise asymptotic power of at least $1-\alpha$, i.e.,*

$$\inf_{\mathbb{P}\in H_{A,S}^{\mathcal{MI}}} \lim_{n\to\infty} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) \geq 1-\alpha.$$

If ϕ_n has uniform asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has uniform asymptotic level and uniform asymptotic power of at least $1-\alpha$.

Due to Proposition 2, a test for $H_{0,S}^{\mathcal{MI}}$ is implicitly a test for $S \subseteq \mathrm{AN}_Y$, and can thus be used to infer whether intervening on S will have a potential causal effect on Y. However, rejecting $H_{0,S}^{\mathcal{MI}}$ is not evidence for $S \subseteq \mathrm{AN}$; it is evidence for $S \notin \mathcal{MI}$.

5.2. Learning S_{IAS} from data

We now consider the task of estimating the set S_{IAS} from data. If we are given a test for invariance that has asymptotic level and power and if we correct for multiple testing appropriately, we can estimate S_{IAS} by $\hat{S}_{\mathrm{IAS}}$, which, asymptotically, is a subset of AN_Y with large probability.

Theorem 2. *Assume that the decision rule ϕ_n has pointwise asymptotic level for all minimally invariant sets and pointwise asymptotic power for all $S \subseteq [d]$ such that S is not a superset of a minimally invariant set. Define $C := 2^d$ and let $\widehat{\mathcal{I}} := \{S \subseteq [d] \mid \phi_n(S, \alpha C^{-1}) = 0)\}$ be the set of all sets for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}} := \left\{S \in \widehat{\mathcal{I}} \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}\right\}$ and $\hat{S}_{\mathrm{IAS}} := \bigcup_{S\in\widehat{\mathcal{MI}}} S$. It then holds that*

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) &\geq \lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}) \\ &\geq 1-\alpha.\end{aligned}$$

A generic algorithm for implementing $\hat{S}_{\mathrm{IAS}}$ is given in Appendix E.4.

Remark 1. Consider a decision rule ϕ_n that just (correctly) rejects the empty set (e.g., because the p-value is just below the threshold α), indicating that the effect of the environments is weak. It is likely that there are other sets $S' \notin \mathcal{I}$, which the test may not have sufficient power against and are (falsely) accepted as invariant. If one of such sets contains non-ancestors of Y, this yields a violation of $\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$. To guard against this, testing $S = \emptyset$ can be done at a lower significance level, $\alpha_0 < \alpha$. This modified IAS approach is conservative and may return $\hat{S}_{\mathrm{IAS}} = \emptyset$ if the environments do not have a strong impact on Y, but it retains the guarantee $\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq 1-\alpha$ of Theorem 2.

The multiple testing correction performed in Theorem 2 is strictly conservative because we only need to correct for the number of minimally invariant sets, and there does not exist 2^d minimally invariant sets. Indeed, the statement of Theorem 2 remains valid for $C = C'$ if the underlying DAG has at most C' minimally invariant sets. We hypothesize that a DAG can contain at most $3^{\lceil d/3 \rceil}$ minimally invariant sets and therefore propose using $C = 3^{\lceil d/3 \rceil}$ in practice. If this hypothesis is true, Theorem 2 remains valid (for any DAG), using $C = 3^{\lceil d/3 \rceil}$ (see Appendix E.3 for a more detailed discussion).

Alternatively, as shown in the following section, we can restrict the search for minimally invariant sets to a predetermined size. This requires milder correction factors and comes with computational benefits.

5.3. Invariant ancestry search in large systems

We now develop a variation of Theorem 2, which allows us to search for ancestors of Y in large graphs, at the cost of only identifying minimally invariant sets up to some a priori determined size.

Similarly to ICP (see Section 2.2), one could restrict IAS to the variables in MB_Y but the output may be smaller than S_{IAS}; in particular, there are only non-parental ancestors in MB_Y if these are parents to both a parent a child of Y (For instance, in the graph $E \to X_1 \to \ldots \to X_d \to Y$, $S_{\mathrm{IAS}} = \{1, \ldots, d\}$ but restricting IAS to MB_Y would yield the set $\{d\}$.) Thus, we do not expect such an approach to be particularly fruitful in learning ancestors.

Here, we propose an alternative approach and define

$$S^m_{\mathrm{IAS}} := \bigcup_{S : S \in \mathcal{MI} \text{ and } |S| \leq m} S \tag{6}$$

as the union of minimally invariant sets that are no larger than $m \leq d$. For computing S^m_{IAS}, one only needs to check invariance of the $\sum_{i=0}^{m} \binom{d}{i}$ sets that are no larger than m. S^m_{IAS} itself, however, can be larger than m: in the graph above (6), $S^1_{\mathrm{IAS}} = \{1, \ldots, d\}$. The following proposition characterizes properties of S^m_{IAS}.

Proposition 5. *Let $m < d$ and let $m_{\min}$ and $m_{\max}$ be the size of a smallest and a largest minimally invariant set, respectively. The following statements are true:*

(i) $S^m_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$.

(ii) If $m \geq m_{\max}$, *then* $S^m_{\mathrm{IAS}} = S_{\mathrm{IAS}}$.

(iii) If $m \geq m_{\min}$ *and* $E \notin \mathrm{PA}_Y$, *then* $S^m_{\mathrm{IAS}} \in \mathcal{I}$.

(iv) If $m \geq m_{\min}$ *and* $E \notin \mathrm{PA}_Y$, *then* $S_{\mathrm{ICP}} \subseteq S^m_{\mathrm{IAS}}$ *with equality if and only if* $S_{\mathrm{ICP}} \in \mathcal{I}$.

If $m < m_{\min}$ and $S_{\mathrm{ICP}} \neq \emptyset$, then $S_{\mathrm{ICP}} \subseteq S^m_{\mathrm{IAS}}$ does not hold. However, we show in Section 6.1 using simulations that S^m_{IAS} is larger than S_{ICP} in many sparse graphs, even for $m = 1$, when few nodes are intervened on.

In addition to the computational speedup offered by considering S^m_{IAS} instead of S_{IAS}, the set S_{IAS} can be estimated from data using a smaller correction factor than the one employed in Theorem 2. This has the benefit that in practice, smaller sample sizes may be needed to detect non-invariance.

Theorem 3. *Let $m \leq d$ and define $C(m) := \sum_{i=0}^{m} \binom{d}{i}$. Assume that the decision rule ϕ_n has pointwise asymptotic level for all minimally invariant sets of size at most m and pointwise power for all sets of size at most m that are not supersets of a minimally invariant set. Let $\widehat{\mathcal{I}}^m := \{S \subseteq [d] \mid \phi_n(S, \alpha C(m)^{-1}) = 0 \text{ and } |S| \leq m\}$, be the set of all sets of size at most m for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}}^m := \left\{S \in \widehat{\mathcal{I}}^m \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}^m\right\}$ and $\hat{S}^m_{\text{IAS}} := \bigcup_{S \in \widehat{\mathcal{MI}}^m} S$. It then holds that*

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\text{IAS}} \subseteq \text{AN}_Y) &\geq \lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\text{IAS}} = S^m_{\text{IAS}}) \\ &\geq 1 - \alpha.\end{aligned}$$

The method proposed in Theorem 3 outputs a non-empty set if there exists a non-empty set of size at most m, for which the hypothesis of invariance cannot be rejected. In a sparse graph, it is likely that many small sets are minimally invariant, whereas if the graph is dense, it may be that all invariant sets are larger than m, such that $S^m_{\text{IAS}} = \emptyset$. In dense graphs however, many other approaches may fail too; for example, it is also likely that the size of the Markov boundary is so large that applying ICP on MB_Y is not feasible.

6. Experiments

We apply the methods developed in this paper in a population-case experiment using oracle knowledge (Section 6.1), a synthetic experiment using finite sample tests (Section 6.2), and a real-world data set from a gene perturbation experiment (Section 6.3). In Sections 6.1 and 6.2 we consider a setting with two environments: an observational environment ($E = 0$) and an intervention environment ($E = 1$), and examine how the strength and number of interventions affect the performance of IAS.

6.1. Oracle IAS in random graphs

For the oracle setting, we know that $S_{\text{IAS}} \subseteq \text{AN}_Y$ (Proposition 2) and $S_{\text{ICP}} \subseteq S_{\text{IAS}}$ (Proposition 3). We first verify that the inclusion $S_{\text{ICP}} \subseteq S_{\text{IAS}}$ is often strict in low-dimensional settings when there are few interventions. Second, we show that the set S^m_{IAS} is often strictly larger than the set $S^{\text{MB}}_{\text{ICP}}$ in large, sparse graphs with few interventions.

In principle, for a given number of covariates, one can enumerate all DAGs and, for each DAG, compare S_{ICP} and S_{IAS}. However, because the space of DAGs grows super-exponentially in the number of nodes [Chickering, 2002], this is infeasible. Instead, we sample graphs from the space of all DAGs that satisfy Assumption 1 and $Y \in \text{DE}_E$ (see Appendix E.5.1 for details).

In the low-dimensional setting ($d \leq 20$), we compute S_{ICP} and S_{IAS}, whereas in the larger graphs ($d \geq 100$), we compute $S_{\text{ICP}}^{\text{MB}}$ and the reduced set S_{IAS}^{m} for $m \in \{1, 2\}$ when $d = 100$ and for $m = 1$ when $d = 1{,}000$. Because there is no guarantee that IAS outputs a superset of ICP when searching only up to sets of some size lower than d, we compare the size of the sets output by either method. For the low-dimensional setting, we consider both sparse and dense graphs, but for larger dimensions, we only consider sparse graphs. In the sparse setting, the DAGs are constructed such that there is an expected number of $d+1$ edges between the $d+1$ nodes X and Y; in the dense setting, the expected number of edges equals $0.75 \cdot d(d+1)/2$.

The results of the simulations are displayed in Figs. 2 and 3. In the low-dimensional setting, S_{IAS} is a strict superset of S_{ICP} for many graphs. This effect is the more pronounced, the larger the d and the fewer nodes are intervened on, see Fig. 2. In fact, when there are interventions on all predictors, we know that $S_{\text{IAS}} = S_{\text{ICP}} = \text{PA}_Y$ [Peters et al., 2016, Theorem 2], and thus the probability that $S_{\text{ICP}} \subsetneq S_{\text{IAS}}$ is exactly zero. For the larger graphs, we find that the set S_{IAS}^{m} is, on average, larger than $S_{\text{ICP}}^{\text{MB}}$, in particular when $d = 1{,}000$ or when $m = 2$, see Fig. 3. In the setting with $d = 100$ and $m = 1$, the two sets are roughly the same size, when 10% of the predictors are intervened on. The set $S_{\text{ICP}}^{\text{MB}}$ becomes larger than S_{IAS}^{1} after roughly 15% of the predictors nodes are intervened on (not shown). For both $d = 100$ and $d = 1{,}000$, the average size of the Markov boundary of Y was found to be approximately 3.5.

6.2. Simulated linear Gaussian SCMs

In this experiment, we show through simulation that IAS finds more ancestors than ICP in a finite sample setting when applied to linear Gaussian SCMs. To compare the outputs of IAS and ICP, we use the *Jaccard similarity* between $\hat{S}_{\text{IAS}}$ ($\hat{S}_{\text{IAS}}^{1}$ when d is large) and AN_Y, and between $\hat{S}_{\text{ICP}}$ ($\hat{S}_{\text{ICP}}^{\widehat{\text{MB}}}$ when d is large[5]) and AN_Y.[6]

We sample data from sparse linear Gaussian models with i.i.d. noise terms in two scenarios, $d = 6$ and $d = 100$. In both cases, coefficients for the linear assignments are drawn randomly. We consider two environments; one observational and one interventional; in the interventional environment, we apply do-interventions of strength one to children of E, i.e., we fix the value of a child of E to be one. We standardize the data along the causal order, to prevent variance accumulation along the causal order [Reisach et al., 2021]. Throughout the section, we consider a significance level of $\alpha = 5\%$. For a detailed description of the simulations, see Appendix E.5.2.

To test for invariance, we employ the test used in Peters et al. [2016]: We calculate a p-value for the hypothesis of invariance of S by first linearly regressing Y onto X_S (ignoring E), and second testing whether the mean and variance of the prediction residuals is equal across environments. For details, see Peters et al. [2016, Section 3.2.1]. Schultheiss et al. [2021] also consider the task of estimating ancestors but since their method is

[5] $\widehat{\text{MB}}$ is a Lasso regression estimate of MB_Y containing at most 10 variables

[6] The Jaccard similarity between two sets A and B is defined as $J(A, B) \coloneqq |A \cap B|/|A \cup B|$, with $J(\emptyset, \emptyset) = 0$. The Jaccard similarity equals one if the two sets are equal, zero if they are disjoint and takes a value in $(0, 1)$ otherwise.

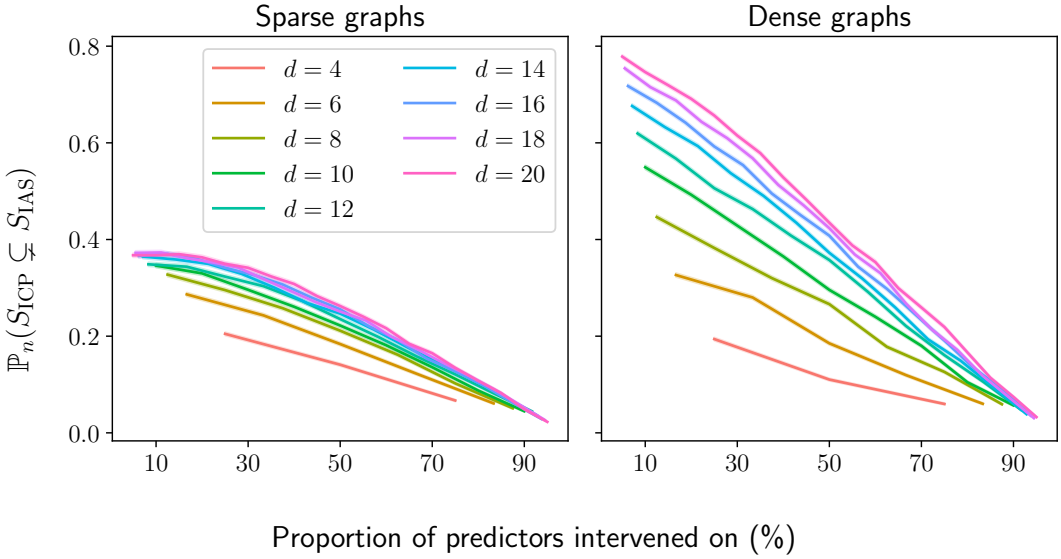


Figure 2.: Low-dimensional oracle experiment, see Section 6.1. In all cases, as predicted by theory, S_{ICP} is contained in S_{IAS}. For many graphs, S_{IAS} is strictly larger than S_{ICP}. On average, this effect is more expressed when there are fewer intervened nodes. $\mathbb{P}_n$ refers to the distribution used to sample graphs and every point in the figure is based on 50,000 independently sampled graphs; d denotes the number of covariates X. Empirical confidence bands are plotted around each line, but are very narrow.

uninformative for Gaussian data and does not consider environments, it is not directly applicable here.

In Theorem 2, we assume asymptotic power of our invariance test. When $d = 6$, we test hypotheses with a correction factor $C = 3^{\lceil 6/3 \rceil} = 9$, as suggested in Appendix E.3, in an attempt to reduce false positive findings. In Appendix E.5.3, we repeat the experiment of this section with $C = 2^6$ and find almost identical results. We hypothesize, that the effects of a reduced C is more pronounced at larger d. When $d = 100$, we test hypotheses with the correction factor $C(1)$ of Theorem 3. In both cases, we test the hypothesis of invariance of the empty set at level $\alpha_0 = 10^{-6}$ (cf. Remark 1). In Appendix E.5.4, we investigate the effects on the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$ and $\mathbb{P}(\hat{S}^1_{\text{IAS}} \subseteq \text{AN}_Y)$ when varying α_0, confirming that choosing α_0 too high can lead to a reduced probability of $\hat{S}_{\text{IAS}}$ being a subset of ancestors.

In Fig. 4 the results of the simulations are displayed. In SCMs where the oracle versions S_{IAS} and S_{ICP} are not equal, $\hat{S}_{\text{IAS}}$ achieved, on average, a higher Jaccard similarity to AN_Y than $\hat{S}_{\text{ICP}}$. This effect is less pronounced when $d = 100$. We believe that the difference in Jaccard similarities is more pronounced when using larger values of m. When $S_{\text{IAS}} = S_{\text{ICP}}$, the two procedures achieve roughly the same Jaccard similarities to AN_Y, as expected. When the number of observations is one hundred, IAS generally

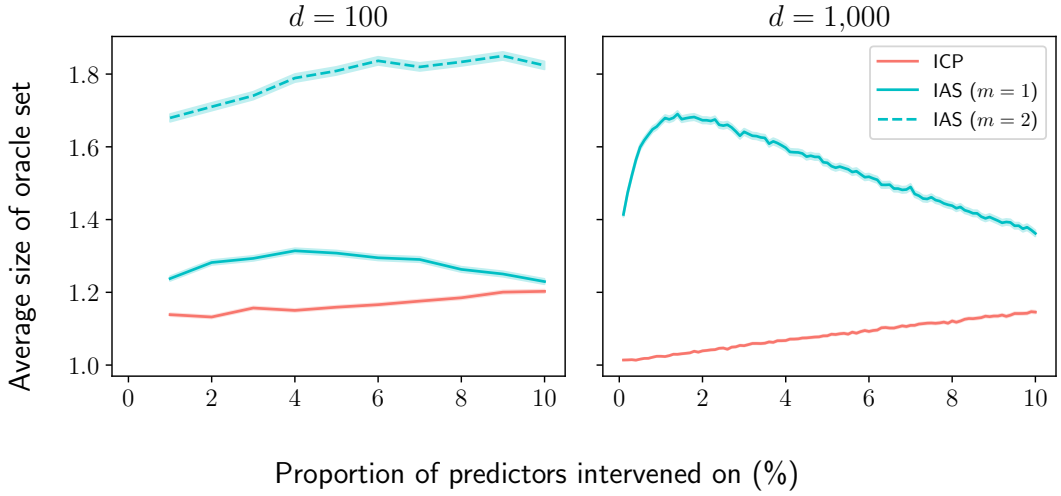


Figure 3.: High-dimensional oracle experiment with sparse graphs, see Section 6.1. The average size of the set S^m_{IAS} is larger than the average size of the set $S^{\text{MB}}_{\text{ICP}}$, both when using IAS to search for sets up to sizes $m = 1$ and $m = 2$. Except for the choice of d, the setup is the same as in Fig. 2.

fails to find any ancestors and outputs the empty set (see Fig. E.2), indicating that the we do not have power to reject the empty set when there are few observations. This is partly by design; we test the empty set for invariance at reduced level α_0 in order to protect against making false positive findings when the environment has a weak effect on Y. However, even without testing the empty set at a reduced level, IAS has to correct for making multiple comparisons, contrary to ICP, thus lowering the marginal significance level each set is tested at. When computing the jaccard similarities with either $\alpha_0 = \alpha$ or $\alpha_0 = 10^{-12}$, the results were similar (not shown). We repeated the experiments with $d = 6$ with a weaker influence of the environment (do-interventions of strength 0.5 instead of 1) and found comparable results, with slightly less power in that the empty set is found more often, see Appendix E.5.5.

We compare our method with a variant, called $\text{IAS}_{\text{est. graph}}$, where we first estimate (e.g., using methods proposed by Mooij et al. 2020 or Squires et al. 2020) a member graph of the Markov equivalence class ('I-MEC') and apply the oracle algorithm from Section 4 (by reading of d-separations in that graph) to estimate $\mathcal{MI}$. In general, however, such an approach comes with additional assumptions; furthermore, even in the linear setup considered here, its empirical performance for large graphs is worse than the proposed method IAS, see Appendix E.5.7.

6.3. IAS in high dimensional genetic data

We evaluate our approach in a data set on gene expression in yeast Kemmeren et al. [2014]. The data contain full-genome mRNA expressions of $d = 6{,}170$ genes and consists

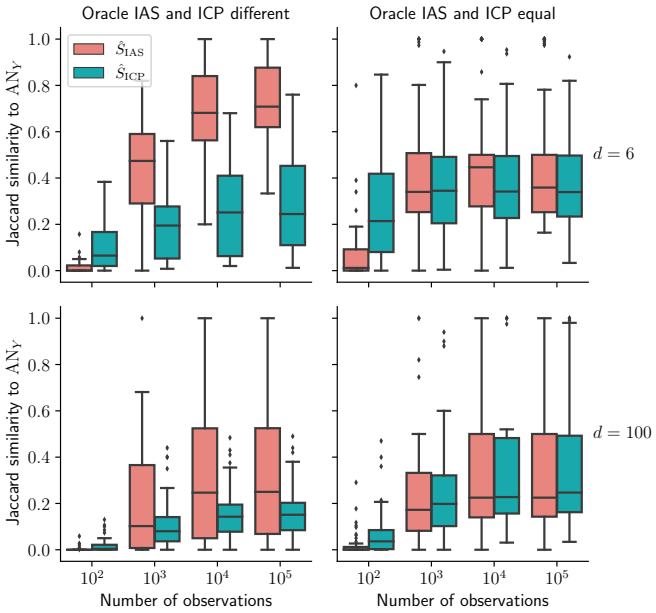


Figure 4.: Comparison between the finite sample output of IAS and ICP and AN_Y on simulated data, see Section 6.2. The plots show the Jaccard similarities between AN_Y and either $\hat{S}_{\text{IAS}}$ ($\hat{S}^1_{\text{IAS}}$ when $d = 100$) in red or $\hat{S}_{\text{ICP}}$ ($\hat{S}^{\widehat{\text{MB}}}_{\text{ICP}}$ when $d = 100$) in blue and AN_Y. When $S_{\text{ICP}} \neq S_{\text{IAS}}$ (left column), $\hat{S}_{\text{IAS}}$ is more similar to AN_Y than $\hat{S}_{\text{ICP}}$. The procedures are roughly equally similar to AN_Y when $S_{\text{ICP}} = S_{\text{IAS}}$ (right column). Graphs represented in each boxplot: 42 (top left), 58 (top right), 40 (bottom left) and 60 (bottom right).

of $n_{\text{obs}} = 160$ unperturbed observations ($E = 0$) and $n_{\text{int}} = 1{,}479$ intervened-upon observations ($E = 1$); each of the latter observations correspond to the deletion of a single (known) gene. For each response gene $\texttt{gene}_Y \in [d]$, we apply the procedure from Section 5.3 with $m = 1$ to search for ancestors.

We first test for invariance of the empty set, i.e., whether the distribution of $\texttt{gene}_Y$ differs between the observational and interventional environment. We test this at a conservative level $\alpha_0 = 10^{-12}$ in order to protect against a high false positive rate (see Remark 1). For 3,631 out of 6,170 response genes, the empty set is invariant, and we disregard them as response genes.

For each response gene, for which the empty set is not invariant, we apply our procedure. More specifically, when testing whether $\texttt{gene}_X$ is an ancestor of $\texttt{gene}_Y$, we exclude any observation in which either $\texttt{gene}_X$ or $\texttt{gene}_Y$ was intervened on. We then test whether the empty set is still rejected, at level $\alpha_0 = 10^{-12}$, and whether $\texttt{gene}_X$ is invariant at level $\alpha = 0.25$. Since a set $\{\texttt{gene}_X\}$ is deemed minimally invariant if the p-value exceeds α, setting α large is conservative for the task of finding ancestors. Indeed, when estimating $\hat{S}^m_{\text{IAS}}$, one can test the sets of size m at a higher level $\alpha_1 > \alpha$. This is conservative, because falsely rejecting a minimally invariant set of size m does not break

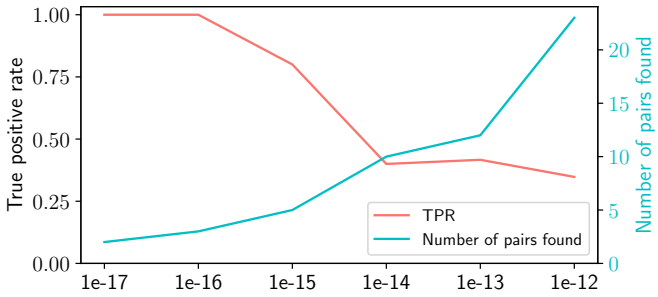


Figure 5.: True positive rates and number of gene pairs found in the experiment in Section 6.3. On the x-axis, we change α_0, the threshold for invariance of the empty set. When α_0 is small, we only search for pairs if the environment has a very significant effect on Y. For smaller α_0, fewer pairs are found to be invariant (blue line), but those found, are more likely to be true positives (red line). This supports the claim that the lower α_0 is, the more conservative our approach is.

the inclusion $\hat{S}^m_{\text{IAS}} \subseteq \text{AN}_Y$. However, if one has little power against the non-invariant sets of size m, testing at level α_1 can protect against false positives.[7]

We use the held-out data point, where $\texttt{gene}_X$ is intervened on, to determine as ground truth, whether $\texttt{gene}_X$ is indeed an ancestor of $\texttt{gene}_Y$. We define $\texttt{gene}_X$ as a true ancestor of $\texttt{gene}_Y$ if the value of $\texttt{gene}_Y$ when $\texttt{gene}_X$ is intervened on, lies in the $q_{TP} = 1\%$ tails of the observational distribution of $\texttt{gene}_Y$.

We find 23 invariant pairs $(\texttt{gene}_X, \texttt{gene}_Y)$; of these, 7 are true positives. In comparison, Peters et al. [2016] applies ICP to the same data, and with the same definition of true positives. They predict 8 pairs, of which 6 are true positives. This difference is in coherence with the motivation put forward in Section 5.2: Our approach predicts many more ancestral pairs (8 for ICP compared to 23 for IAS). Since ICP does not depend on power of the test, they have a lower false positive rate (25% for ICP compared to 69.6% for IAS).

In Fig. 5, we explore how changing α_0 and q_{TP} impacts the true positive rate. Reducing α_0 increases the true positive rate, but lowers the number of gene pairs found (see Fig. 5). This is because a lower α_0 makes it more difficult to detect non-invariance of the empty set, making the procedure more conservative (with respect to finding ancestors); see Remark 1. For example, when $\alpha_0 \leq 10^{-15}$, the true positive rate is above 0.8; however, 5 or fewer pairs are found. When searching for ancestors, the effect of intervening may be reduced by noise from intermediary variables, so $q_{TB} = 1\%$ might be too strict; in Appendix E.5.6, we analyze the impact of increasing q_{TB}.

[7] Only sets of size exactly m can be tested at level α_1; the remaining hypotheses should still be corrected by $C(m)$ (or by the hypothesized number of minimally invariant sets).

7. Extensions

7.1. Latent variables

In Assumption 1, we assume that all variables X are observed and that there are no hidden variables H. Let us write $X = X_O \dot{\cup} X_H$, where only X_O is observed and define $\mathcal{I} \coloneqq \{S \subseteq X_O \mid S \text{ invariant}\}$. We can then define

$$S_{\mathrm{IAS},O} \coloneqq \bigcup_{S \subseteq X_O : H_{0,S}^{\mathcal{M}^{\mathcal{I}}} \text{ true}} S$$

(again with the convention that a union over the empty set is the empty set), and have the following modification of Proposition 2.

Proposition 6. *It holds that* $S_{\mathrm{IAS},O} \subseteq \mathrm{AN}_Y$.

All results in this paper remain correct in the presence of hidden variables, except for Proposition 3 and Proposition 5 (iii-iv).[8] Thus, the union of the observed minimally invariant sets, $S_{\mathrm{IAS},O}$ is a subset of AN_Y and can be learned from data in the same way as if no latent variables were present.

7.2. Non-exogenous environments

Throughout this paper, we have assumed that the environment variable is exogenous (Assumption 1). However, all of the results stated in this paper, except for Proposition 4, also hold under the alternative assumption that E is an ancestor of Y, but not necessarily exogenous. From the remaining results, only the proof of Proposition 1 uses exogeneity of E, but here the result follows from Tian et al. [1998]. In all other proofs, we account for both options. This extension also remains valid in the presence of hidden variables, using the same arguments as in Section 7.1.

8. Conclusion and Future Work

Invariant Ancestry Search (IAS) provides a framework for searching for causal ancestors of a response variable Y through finding minimally invariant sets of predictors by exploiting the existence of exogenous heterogeneity. The set S_{IAS} is a subset of the ancestors of Y, a superset of S_{ICP} and, contrary to S_{ICP}, invariant itself. Furthermore, the hierarchical structure of minimally invariant sets allows IAS to search for causal ancestors only among subsets up to a predetermined size. This avoids exponential runtime and allows us to apply the algorithm to large systems. We have shown that, asymptotically, S_{IAS} can be identified from data with high probability if we are provided with a test

[8] These results do not hold in the presence of hidden variables, because it is not guaranteed that an invariant set exists among X_O (e.g., consider a graph where all observed variables share a common, unobserved confounder with Y). However, if at least one minimally invariant set exists among the observed variables, then all results stated in this paper hold.

for invariance that has asymptotic level and power. We have validated our procedure both on simulated and real data. Our proposed framework would benefit from further research in the maximal number of minimally invariant sets among graphs of a fixed size, as this would provide larger finite sample power for identifying ancestors. Further it is of interest to establish finite sample guarantees or convergence rates for IAS, possibly by imposing additional assumptions on the class of SCMs. Finally, even though current implementations are fast, it is an open theoretical question whether computing S_{IAS} in the oracle setting of Section 4 is NP-hard, see Appendix E.2.

Acknowledgements

NT and JP were supported by a research grant (18968) from VILLUM FONDEN.

Identifying Causal Effects using Instrumental Time Series: Nuisance IV and Correcting for the Past

Nikolaj Thams, Rikke Søndergaard, Sebastian Weichwald and Jonas Peters

Abstract

Instrumental variable (IV) regression relies on instruments to infer causal effects from observational data with unobserved confounding. We consider IV regression in time series models, such as vector auto-regressive (VAR) processes. Direct applications of i.i.d. techniques are generally inconsistent as they do not correctly adjust for dependencies in the past. In this paper, we propose methodology for constructing identifying equations that can be used for consistently estimating causal effects. To do so, we develop nuisance IV, which can be of interest even in the i.i.d. case, as it generalizes existing IV methods. We further propose a graph marginalization framework that allows us to apply nuisance and other IV methods in a principled way to time series. Our framework builds on the global Markov property, which we prove holds for VAR processes. For VAR(1) processes, we prove identifiability conditions that relate to Jordan forms and are different from the well-known rank conditions in the i.i.d. case (they do not require as many instruments as covariates, for example). We provide methods, prove their consistency, and show how the inferred causal effect can be used for distribution generalization. Simulation experiments corroborate our theoretical results. We provide ready-to-use Python code.

1. Introduction

Predicting a response variable Y from observations of covariates X may be insufficient to answer a scientific question at hand. Instead, we may wish to model how the response variable Y reacts to an intervention on X. Such modeling requires causal knowledge. For example, for i.i.d. data from a linear model $Y := \beta X + g(H, \varepsilon^Y)$, it is well-known that an ordinary least squares (OLS) regression of Y on X generally yields a biased estimator of the linear causal effect β from X on Y when an unobserved effect H confounds X and Y. Instead, we may obtain unbiased estimates of β by utilising instrumental variables

(IVs) I that correlate with the covariates X, are independent of H, and affect Y only indirectly through X. IV regression, pioneered by Wright [1928] and Reiersøl [1945], is well-established in econometrics [Angrist et al., 1996, Staiger and Stock, 1997, Angrist and Krueger, 2001], statistics [Bowden and Turkington, 1985] and epidemiology [Hernán and Robins, 2006, Didelez et al., 2010]. One approach for IV estimation in the linear i.i.d. model is the two-stage least squares (TSLS) estimator [Angrist and Imbens, 1995], which first estimates the effect from I to X (stage 1) and then regresses Y on the fitted values from the first regression (stage 2). Another formulation, used by Hansen [1982], is the generalized method of moments (GMM), which uses the independence of the residual $Y - \beta X = g(H, \varepsilon^Y)$ from the instrument I: One can estimate β by selecting $\hat{\beta}$ such that the empirical correlation between $Y - \hat{\beta}X$ and I is minimized. If the dimension of I is greater than or equal to the dimension of X, these estimators are consistent [e.g., Hall, 2005].

In more recent approaches, causality and directed acyclic graph (DAG) representations have proved fruitful for studying instrumental variables for i.i.d. data [Pearl, 2009, Hernán and Robins, 2006, Didelez et al., 2010]. Brito and Pearl [2002a] proposed 'generalized IV', a graphical framework that enlarges the class of graphical models, in which IV methods can be used to identify causal effects. Similarly, 'conditional IV' [Pearl, 2009] relaxes the assumptions of IV by considering a conditional moment equation [see also Henckel, 2021].

In many real-world applications (see Weigend [2018] for examples from various fields), the data are sampled not independently but rather as a time series that exhibits memory effects, with past values affecting present ones. However, using IV methods in time series data poses a number of challenges. For example, memory effects in the observed processes X, Y and I can obfuscate the assumption that I only affects present values of Y through the present value of X, because I and Y are confounded by common ancestors in the memory of the process. Additionally, memory effects, or serially correlated errors, in the confounder process H can make identification of the dependence on past states of the process difficult; for such settings, Fair [1970] proposes a search-based method for a subclass of first order vector auto-regressive (VAR) processes. If one is provided with identifying equations with serially correlated errors (such as the ones proposed in this paper), Newey and West [1987] construct confidence intervals by using heteroskedasticity and auto-correlation consistent (HAC) estimators to estimate long-run covariance matrices.

Graphical models have been studied (also in the context of causal inference) when data follow a time series structure [e.g., Wiener, 1956, Granger, 1969, 1980, Dahlhaus and Eichler, 2003, Didelez, 2008, Danks and Plis, 2013, Hyttinen et al., 2016, Peters et al., 2017, Mogensen and Hansen, 2020] but not in the context of instrumental variables and hidden confounders. In this work, we establish a link between graphical models and IV methods for time series, which we then exploit to construct estimators and prove consistency. To help build this connection, we prove (Theorem 1) that the global Markov property [Lauritzen, 1996] holds in VAR(p) processes. To the best of our knowledge this result has not been proved before and requires, due to the graph containing infinitely many nodes, technically involved arguments.

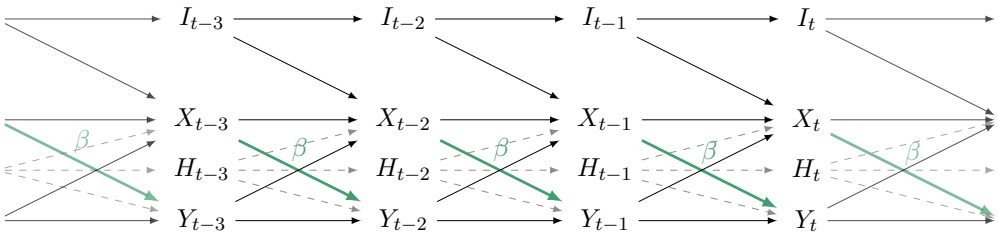


Figure 1.: Full time graph (formally defined in Section 2.2) of a process S satisfying the conditions studied in this paper, see Assumption (A2) below. We aim to estimate the causal effect β (highlighted in green) of X_{t-1} on Y_t, where $X = [X_t]_{t\in\mathbb{Z}}$ and $Y = [Y_t]_{t\in\mathbb{Z}}$ are subprocesses of S, that are confounded by a latent process $H = [H_t]_{t\in\mathbb{Z}}$. Motivated by instrumental variables, one may aim to exploit the subprocess $I = [I_t]_{t\in\mathbb{Z}}$ that is independent of H and only acts on Y through X. However, simply using I_{t-2} as an instrument is generally inconsistent; the same holds when adding X_{t-2} and Y_{t-1} as a conditioning set, for example. This paper develops a graphical framework giving rise to several consistent estimators.

Throughout this work we consider a joint process $S := \left[I_t^\top, X_t^\top, H_t^\top, Y_t^\top\right]_{t\in\mathbb{Z}}^\top$ where H is latent. For simplicity, we let S be a linear VAR(1) process but we show later (see Section 4.2.3) that the methodology applies to broader model classes, too. Fig. 1 shows a graphical representation of the process. Our goal is to estimate the coefficient β with which X_{t-1} linearly enters into Y_t . The coefficient β can also be understood as the direct causal effect of X_{t-1} on Y_t (cf. Appendix F.1.3).

When S is fully observed, estimators that are consistent and asymptotically normal for the standard form parameters of the VAR(1) process S exist; in particular, β can be consistently estimated [Hamilton, 1994]. Yet in our setting H, the confounding between X and Y is unobserved, and such estimators are not applicable. Furthermore, a direct application of IV estimation for i.i.d. data could suggest using I_{t-2} as an instrument for estimating the effect of X_{t-1} on Y_t. We prove that in general, this does not yield a consistent estimator (Proposition 2). The reason is that the instruments are correlated to Y not only through the path $I_{t-2} \to X_{t-1} \to Y_t$ but also through an infinite number of paths in the past, due to common ancestors $I_{t-j}, j \geq 3$ in the instrument process I. This correlation violates the assumption that the instrument I_{t-2} only correlates with Y_t through X_{t-1}.

In this work, we establish a general graph marginalization technique that allows us to find valid IV models for processes S, such as the one shown in Fig. 1. Based on these results, we propose two solutions to identify causal effects in time series. The first solution ('conditional IV' or 'CIV') identifies β using IV conditioned on one or more past states $I_{t-j}, j \geq 3$ of the instrument process. The second solution is based on 'nuisance IV' (or 'NIV'), a modification of IV that we develop; it can be used not only for time series but also for i.i.d. data. It allows for stronger identifiability results by adding nuisance treatment variables to the target causal effect. Applied to the time

series setting, nuisance IV yields a consistent estimator for the target of inference β by adding the effect of one or more nuisance regressors, e.g., Y_{t-1}.

Similar to the i.i.d. case, these two approaches induce identifying moment equations that are satisfied by the causal effect β. Rank conditions guarantee that their solution is unique, allowing us to identify the causal effect. Unlike in the i.i.d. case, however, the standard conditions are not easily interpretable in the time series setting. We therefore develop sufficient and necessary conditions on the parameters of the data-generating process that provide insight on when identifiability holds. Our results imply that identifiability with nuisance IV depends on geometric multiplicities of eigenvalues of the parameter matrix in the VAR process, and we show that if parameters are drawn at random from a continuous distribution, the causal effect β of X_{t-1} on Y_t is almost surely identifiable. In particular, it is possible to identify the causal effect even if the instrument I is univariate and the regressor X is multivariate. For both of the approaches (conditional IV and nuisance IV), we propose estimators and prove that, in case of identifiability, these estimators consistently estimate the direct causal effect.

Finally, we apply our findings to the task of distribution generalization [e.g., Meinshausen, 2018, Rothenhäusler et al., 2021, Jakobsen and Peters, 2021]. In many systems, the causal effects are of value in themselves because they contribute to the understanding of the system but it also serves a purpose when predicting Y_{t+1} under an intervention on X_t. In a linear setting, the OLS estimator has the smallest expected mean squared error (MSE) among all linear predictors when predicting new test data from the observational distribution. However, as is known for the i.i.d. setting [e.g., Rojas-Carulla et al., 2018], causal estimators have better worst-case predictive performance when the environments are constructed by interventions on the covariates. Similarly, we show that in time series, under arbitrary interventions on X_t, our IV estimators are worst-case prediction optimal for Y_{t+1}.

Our work is structured as follows. Section 2 introduces the model and the assumptions considered in this paper; we review graphical representations of time series models and prove that the global Markov property holds for VAR(p) processes. In Section 3, we review theory on conditional instrumental variables, introduce the concept of nuisance IV, and prove its correctness. Our main results for instrumental variable regression for time series are presented in Section 4. We propose two approaches to overcome confounding from past values yielding identifying equations for the causal effect: the first one is based on CIV and the second one uses NIV. For the latter, we characterize identifiability of the causal parameter in terms of parameters of the data-generating process. We also discuss how to use the causal effect to perform optimal prediction of Y_{t+1} under interventions on X_t. In Section 5 we empirically evaluate our method. All proofs are provided in Appendix F.4. Code can be found at `https://github.com/nikolajthams/its-time`.

2. Causal Time Series Models with Confounding

2.1. Definitions and notation

We consider multivariate time series $X := [X_t]_{t\in\mathbb{Z}}$ and $I := [I_t]_{t\in\mathbb{Z}}$, a univariate process $Y := [Y_t]_{t\in\mathbb{Z}}$, and an unobserved multivariate process $H := [H_t]_{t\in\mathbb{Z}}$. Let d_X be the dimensionality of X_t, that is $X_t \in \mathbb{R}^{d_X}$, and similarly for d_I, d_Y and d_H, with $d_Y = 1$. Let $S := [S_t]_{t\in\mathbb{Z}} = [I_t^\top, H_t^\top, X_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$, $S_t \in \mathbb{R}^d$, with $d := d_X + d_Y + d_I + d_H$.

Several of our results are presented for VAR(p) processes [e.g., Brockwell and Davis, 1991], which is done for presentation purposes. Many of the results hold more generally, see Section 4.2.3. We say that S is a VAR(p) process if there exist $p \in \mathbb{N}$ such that the following assumption holds:

(A1) There are coefficient matrices $A_1, \ldots, A_p \in \mathbb{R}^{d\times d}$ such that for all $t \in \mathbb{Z}$:

$$S_t = A_1 S_{t-1} + \ldots + A_p S_{t-p} + \varepsilon_t, \tag{1}$$

where $A_1, \ldots, A_p$ are such that $\det(I_d\lambda^p - A_1\lambda^{p-1} - A_2\lambda^{p-2} - \ldots - A_p) = 0$ implies $|\lambda| < 1$, the ε_t constitute an i.i.d. process, and $\varepsilon_t \sim \mathcal{N}(0, \Gamma)$, where Γ is a diagonal matrix.

We use the notation $\alpha^1_{X,I}$ to refer to the submatrix of A_1 with rows corresponding to X and columns corresponding to I (see Fig. 2a for an example), and similarly $\alpha^2_{I,I}$ etc. We use superscripts to denote individual components of ε, e.g., ε^Y. We consider Y_t as the *response variable*, X_{t-1} as *covariates* and our target of inference is $\beta := \alpha^1_{Y,X}$. We refer to β as the *causal effect* from X to Y.[9] Most of our results generalize to estimating total causal effects (TCEs) instead of β; we state Theorem 3 below in this generality, but for simplicity we state all other results with the causal effect β. For $1 \le i, j \le d$ and $l \in \mathbb{N}$, the TCE of S^i_{t-l} on S^j_t is defined as

$$\left(\sum_{\substack{1\le l_1,\ldots,l_m\le p \\ l_1+\cdots+l_m=l}} A_{l_1} \cdots A_{l_m} \right)_{j,i}.$$

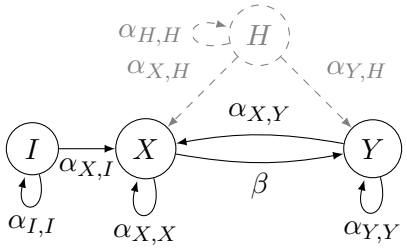
In a process satisfying Assumption (A1), the TCE of X_{t-1} on Y_t coincides with β; see Appendix F.1.4 for details.

Both H and the noise ε are assumed to be unobserved; while the sequence of innovations ε^Y_t is assumed to be i.i.d. and independent of X_t, H can act as a confounder between X and Y and can have an autoregressive structure. Similar to the i.i.d. case [e.g., Hernán and Robins, 2006, Pearl, 2009, Peters et al., 2017], the existence of the confounder H implies that we cannot identify β by simply regressing Y_t on X_{t-1}. In

[9] The notions of causal effect and total causal effect are motivated by interpreting the VAR equations as a structural causal model (SCM), which we explain in detail in Appendix F.1.3. The interventional interpretation of an SCM is not required for any results of the paper, except for the ones presented in Section 4.3, where we discuss optimal predictions under interventions.

$$A_1 = \begin{array}{c} \\ \text{I} \\ \text{H} \\ \text{X} \\ \text{Y} \end{array} \begin{pmatrix} \alpha_{I,I} & 0 & 0 & 0 \\ 0 & \alpha_{H,H} & 0 & 0 \\ \alpha_{X,I} & \alpha_{X,H} & \alpha_{X,X} & \alpha_{X,Y} \\ 0 & \alpha_{Y,H} & \beta & \alpha_{Y,Y} \end{pmatrix}$$

(a) Matrix block structure of A_1 assumed in Assumption (A2).



(b) Summary time graph of a VAR(1) process satisfying Assumption (A2).

Figure 2.: The sparsity structure on the parameter matrix A_1 assumed in Assumption (A2), and a representation of the graphical structure induced by A_1. Zeros in panel a) correspond to absent edges in panel b).

Appendix F.1.2, we provide an example of two VAR(1) processes with two different parameter matrices that generate the same distribution over the observed time series.

We assume that the process has zero mean[10], so for instance $\text{cov}(X_t, Y_t) = \mathbb{E}\{X_t Y_t^\top\} \in \mathbb{R}^{d_X \times d_Y}$. We assume that the data are sampled as follows: We obtain a sample over time points $t = 1, \dots, T$ of S such that for $t = 1$, S_1 follows the stationary distribution. We denote the sample with boldface $\mathbf{S} = [\mathbf{S}_t]_{t=1}^T$ such that $\mathbf{S} \in \mathbb{R}^{d \times T}$ and each column $\mathbf{S}_t$ represents the process observation at time t. We let $\hat{\mathbb{E}}\mathbf{S}_t := \frac{1}{T}\sum_{t=1}^T \mathbf{S}_t$ denote the empirical mean of the distribution (here, the index t in $\hat{\mathbb{E}}\mathbf{S}_t$ does not refer to any specific time point). From Assumption (A1) it follows [Hamilton, 1994, Chapter 10] that

$$\hat{\mathbb{E}}\mathbf{S}_t \xrightarrow{P} \mathbb{E}S_t \quad \text{and} \quad \hat{\text{cov}}\{\mathbf{S}_{t-j}\mathbf{S}_t^\top\} := \frac{1}{T-j}\sum_{t=j+1}^{T} \mathbf{S}_{t-j}\mathbf{S}_t^\top \xrightarrow{P} \text{cov}\{S_{t-j}S_t^\top\}. \tag{2}$$

Finally, in the case where S is a VAR(1) process, we will sometimes assume additional structure on the coefficient matrix.

(A2) Assume that S satisfies Assumption (A1) for $p = 1$ and that A_1 has the sparsity structure displayed in Fig. 2a.

Under Assumption (A2), none of the other time series components enters the assignment for I; in that case, we refer to I as an *instrumental time series*.

2.2. Graph representations of VAR processes

(1) can be represented graphically. This representation will prove helpful when establishing identifying equations for causal effects and constructing consistent estimators. The *full time graph* [e.g., Peters et al., 2013] is defined as an infinite directed graph with nodes I_t, H_t, X_t, Y_t, for any $t \in \mathbb{Z}$. For $k \in \mathbb{N}$, it contains a directed edge from (j, t) to

[10] Since we can always subtract empirical means, the assumption of vanishing means does not come with any loss of generality.

$(i, t+k)$, $i, j \in \{I, H, X, Y\}$, if $\alpha_{ij}^k \neq 0$. For a process satisfying Assumption (A2) an extract of this graph is shown in Fig. 1. We define the full time graph for higher order VAR processes accordingly. The *summary time graph* only has a single node per time series component. It contains a directed edge from i to j, for some $i, j \in \{I, H, X, Y\}$ if and only if the full time graph contains an edge from (j, t) to $(i, t+k)$ for some $k \in \mathbb{N}$. For a process satisfying Assumption (A2), such a graph is visualized in Fig. 2b.

We now introduce some standard graph terminology [e.g., Lauritzen, 1996, Koller and Friedman, 2009, Pearl, 2009]. A *path* p, $p = (v_1, e_1, v_2, \ldots, e_{n-1}, v_n)$, is an alternating sequence of distinct vertices v_i and edges e_i such that v_i and v_{i+1} are connected by e_i. We say that p is a *directed path* from v_1 to v_n if for every i, e_i points from v_i to v_{i+1}. For two nodes v and u, we say that u is a *descendant* of v if there exists a directed path from v to u and otherwise u is a *non-descendant* of v. We write $\mathrm{ND}(v)$ and $\mathrm{DE}\, v$ for the sets of non-descendants and descendants of v, respectively, using the convention that neither of them contain v itself. For a path p and any $i \in \{2, \ldots, n-1\}$, we say that v_i is a *collider on* p if $(v_{i-1}, e_{i-1}, v_i, e_i, v_{i+1})$ is of the form $v_{i-1} \to v_i \leftarrow v_{i+1}$ and else v_i is a *non-collider on* p. We say that the path p is *unblocked, given the set* B if for every non-collider v_i in p, $v_i \notin B$ and for every collider v_i on p, $(v_i \cup \mathrm{DE}\, v_i) \cap B \neq \emptyset$. Otherwise, we say that p is *blocked* by B. If all paths between distinct vertices v and u are blocked by a set B neither containing v nor u, we say that v and u are *d-separated* by B. Similarly, we say that disjoint sets V and U are d-separated by B if all nodes $v \in V$ and $u \in U$ are d-separated by B.

In section Section 4.2, we also consider marginalized graphs, which are acylic directed mixed graphs (ADMGs), containing both directed ($\to$) and bidirected ($\leftrightarrow$) edges. If we define v to be a collider on a path whenever two surrounding edges have arrowheads at v (e.g. $u_1 \leftrightarrow v \leftarrow u_2$), and define descendants only with respect to directed edges, d-separation also extends to ADMGs. See Richardson [2003] for details.

2.2.1. Markov Properties of VAR processes

The representation described above satisfies several Markov properties, which enables us to read off conditional independences from the full time graph. This will be an important tool in our theory, because it enables the use of graphical models to develop IV methodology in time series. These results do not formally follow from standard results in for example Lauritzen [1996], as the full time graphs are infinite, but will be used in many of the proofs.

Theorem 1. *Consider a time series S generated according to Assumption (A1), and finite disjoint collections A, B, C. If A and C are d-separated given B in $\mathcal{G}_{full}$ then $A \perp\!\!\!\perp C | B$.*

The proof of Theorem 1 and all other proofs in this paper are provided in Appendix F.4.

3. Nuisance Effects in Instrumental Variable Regression

In this work, we establish two identifying equations for causal effect estimation in time series: The first one is based on conditional instrumental variables (CIV) and the second one on a generalization that we term nuisance instrumental variables (NIV). We regard the idea of NIV as interesting in its own right, as it can be applied in the i.i.d. setting, too. In this section, we therefore first review CIV regression for i.i.d. data, and then introduce NIV regression; instrumental variable (IV) regression is a special case of CIV regression, where the conditioning set is empty. In Section 4, we extend the CIV and NIV estimators to VAR processes via a reduction of the full time graph to a marginalized graph.

3.1. Instrumental variables and conditional instrumental variables

Consider a linear SCM (see Appendix F.1.3) over variables V, and let $\mathcal{I}, \mathcal{X}, \mathcal{B}, \{Y\} \subseteq V$ be disjoint collections of variables[11] from V, and let $\mathcal{G}$ be the corresponding DAG. Assume that $\mathcal{I}, \mathcal{X}$ and Y have zero mean and finite second moment and let β be the causal coefficient with which $\mathcal{X}$ enters the structural equation for Y, that is,

$$Y = \beta\mathcal{X} + \gamma W + \varepsilon^Y,$$

for some variables $W \subseteq V \setminus \mathcal{X}$; (some of the entries of β can be zero, so not all variables in $\mathcal{X}$ have to be parents of Y). We consider the following three requirements on $\mathcal{I}, \mathcal{X}, \mathcal{B}$ and Y:

(CIV1) $\mathcal{I}$ and Y are d-separated given $\mathcal{B}$ in the graph $\mathcal{G}_{\mathcal{X}\not\to Y}$, that is the graph $\mathcal{G}$ where all direct edges from $\mathcal{X}$ to Y are removed,

(CIV2) $\mathcal{B}$ is not a descendant of $\mathcal{X} \cup Y$ in $\mathcal{G}$, and

(CIV3) the matrix $\mathbb{E}[\mathrm{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})]$ has rank $d_{\mathcal{X}}$, that is, full row rank.

If requirements (CIV1) and (CIV2) are met, $Y - \beta\mathcal{X} \perp\!\!\!\perp \mathcal{I}|\mathcal{B}$, and in particular β satisfies the *CIV moment equation*[12]

$$\mathbb{E}[\mathrm{cov}(Y - \beta\mathcal{X}, \mathcal{I}|\mathcal{B})] = 0. \tag{3}$$

If, additionally, requirement (CIV3) is met, β is the unique solution to (3),

$$\mathbb{E}[\mathrm{cov}(Y - b\mathcal{X}, \mathcal{I}|\mathcal{B})] = 0 \implies b = \beta.$$

[11] Below, the different variables will take different roles (such as instruments or regressors). We use the calligraphic notation $\mathcal{I}, \mathcal{X}$, and $\mathcal{B}$ to denote collections of observed variables, being used as instruments, regressors, and conditioning sets, respectively. Individual variables are denoted by non-calligraphic letters, such as I.

[12] Here we use the definition $\mathrm{cov}(A, C|B) := \mathbb{E}[AC^\top|B] - \mathbb{E}[A|B]\mathbb{E}[C^\top|B] = \mathrm{cov}(A - \mathbb{E}[A|B], C - \mathbb{E}[C|B])$, which even accommodates for nonlinear relationships between the variables and $\mathcal{B}$.

(Conditional IV with univariate $\mathcal{X}$ has been discussed in the literature [Pearl, 2009, Henckel, 2021, Brito and Pearl, 2002a]. Since we allow $d_\mathcal{X} > 1$, we add a short proof in Appendix F.4.1.) In this case, we say that β is *identified by CIV* or, more precisely, *identified by* $\mathrm{CIV}_{\mathcal{X}\to Y}(\mathcal{I}|\mathcal{B})$. If requirements (CIV1) and (CIV3) are satisfied for $\mathcal{B} = \emptyset$ (requirement (CIV2) is trivially satisfied for $\mathcal{B} = \emptyset$), $\mathrm{CIV}_{\mathcal{X}\to Y}(\mathcal{I}|\emptyset)$ reduces to instrumental variables (IV) regression [Reiersøl, 1945, Anderson and Rubin, 1949, Bowden and Turkington, 1985, Angrist et al., 1996], which we refer to as $\mathrm{IV}_{\mathcal{X}\to Y}(\mathcal{I})$. We use the term *proper* CIV when $\mathcal{B} \neq \emptyset$.

For a finite sample $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{I}$, and $\mathbf{B}$, we consider an empirical counterpart of (3) which, however, may not have a solution in the overidentified setting, that is when $d_\mathcal{I} > d_\mathcal{X}$; to overcome this, for any positive definite weight matrix W, we define the estimator $\hat{b}(W)$ as

$$\hat{b}(W) := \arg\min_b \|\hat{\mathrm{cov}}(\mathbf{Y} - b\,\mathbf{X}, \mathbf{I}|\mathbf{B})\|_W^2, \tag{4}$$

where $\|x\|_W^2 := x^\top W x$ and $\hat{\mathrm{cov}}$ is the empirical covariance of the residuals after regressing out $\mathbf{B}$. We refer to this estimator as $\mathrm{CIV}_{\mathbf{X}\to\mathbf{Y}}(\mathbf{I}|\mathbf{B})$. If $\mathcal{I}, \mathcal{X}, Y$ and $\mathcal{B}$ are zero mean random vectors, the minimizer of (4) is given by

$$\hat{b}(W) = \hat{\mathbb{E}}[r_\mathbf{Y} r_\mathbf{I}^\top]\; W\; \hat{\mathbb{E}}[r_\mathbf{I} r_\mathbf{X}^\top] \left(\hat{\mathbb{E}}[r_\mathbf{X} r_\mathbf{I}^\top]\; W\; \hat{\mathbb{E}}[r_\mathbf{I} r_\mathbf{X}^\top] \right)^{-1}, \tag{5}$$

where $r_\mathbf{Y} := \mathbf{Y} - \hat{\mathbb{E}}[\mathbf{Y}|\mathbf{B}]$ are the residuals after regressing $\mathbf{Y}$ on $\mathbf{B}$, and similarly $r_\mathbf{X} := \mathbf{X} - \hat{\mathbb{E}}[\mathbf{X}|\mathbf{B}]$ and $r_\mathbf{I} := \mathbf{I} - \hat{\mathbb{E}}[\mathbf{I}|\mathbf{B}]$.

Choosing the two-stage least squares (TSLS) weight matrix $W_\mathrm{TSLS} := \mathbb{E}[r_\mathbf{I} r_\mathbf{I}^\top]^{-1}$ corresponds to the procedure where one regresses $r_\mathbf{X}$ on $r_\mathbf{I}$ and then returns the regression coefficient of $r_\mathbf{Y}$ on the fitted values $\hat{r}_\mathbf{X}$. In a linear Gaussian model, the IV estimator $\hat{b}(W_\mathrm{TSLS})$ has the lowest asymptotic variance among all positive definite weight matrices W [Hall, 2005].

3.2. Nuisance instrumental variables

IV estimation is a special case of CIV with the empty set as conditioning set. The example in Fig. 3 (left) shows a graph where an effect between X and Y cannot be identified using the IV estimator: no variable is d-separated from H, and hence no valid instruments for (unconditional) IV exist. Yet, the causal effect is identified by[13] $\mathrm{CIV}_{X\to Y}(I|B)$ because B satisfies requirements (CIV1) to (CIV3).

In the case shown in Fig. 3 (middle), the effect from X to Y cannot be identified by $\mathrm{IV}_{X\to Y}(I)$ because of the unblocked path $I \to Z \to Y$. We cannot use proper CIV, either, because the path $I \to Z \leftarrow H \to Y$ is unblocked given Z, violating requirement (CIV1). Nevertheless, we can identify the effect from X to Y by adding an additional regressor variable. If $d_I \geq d_X + d_Z$, then $\mathrm{IV}_{\{X,Z\}\to Y}(I)$ satisfies the

[13] In a slight abuse of notation, we sometimes omit parantheses indicating sets and write $\mathrm{CIV}_{\mathcal{X}\to Y}(I|\mathcal{B})$ instead of $\mathrm{CIV}_{\mathcal{X}\to Y}(\mathcal{I}|\mathcal{B})$ if $\mathcal{I} = \{I\}$, for example.

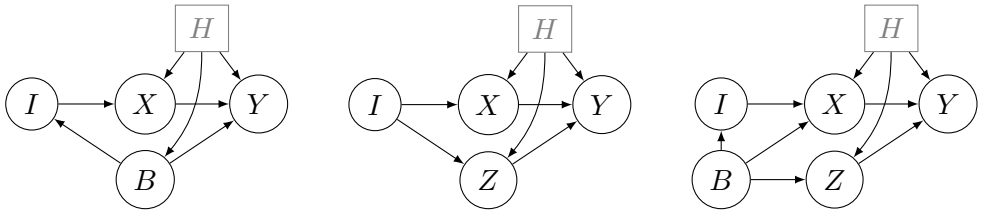


Figure 3.: CIV and (proper) NIV are complementary for identifying causal effects in that they can be used in different settings. (*Left*) A graph where the effect $X \to Y$ is identified by $\mathrm{CIV}_{X\to Y}(I|B)$ (provided that requirement (CIV3) holds) but not by any IV or proper NIV method. (*Middle*) A graph where $X \to Y$ is identified by $\mathrm{NIV}_{X\to Y}(I, Z)$, but not by any proper CIV method. (*Right*) A graph where $X \to Y$ is identified by both $\mathrm{CIV}_{X\to Y}(I|B)$ and $\mathrm{NIV}_{X\to Y}(\mathcal{I}, Z)$ with $\mathcal{I} = \{I, B\}$.

assumption for identifying the effect $\{X, Z\} \to Y$. In particular, from this we can extract the effect of interest, $X \to Y$.

We refer to this approach as *nuisance instrumental variables (NIV)*. More formally, consider collections of variables $\mathcal{I}, \mathcal{X}, \mathcal{Z}$ and a response variable Y. We say that β satisfies the *NIV moment equation* if there exist $\alpha \in \mathbb{R}^{d_Y \times d_\mathcal{Z}}$ such that

$$\mathrm{cov}(Y - \beta\mathcal{X} - \alpha\mathcal{Z}, \mathcal{I}) = 0. \tag{6}$$

We say that β is *identified by NIV* or, more formally *identified by* $\mathrm{NIV}_{\mathcal{X}\to\mathcal{Y}}(I, Z)$ if additionally β is the only solution to the moment equation; that is for all $a \in \mathbb{R}^{d_Y \times d_\mathcal{Z}}$ and $b \in \mathbb{R}^{d_Y \times d_\mathcal{X}}$

$$\mathrm{cov}(Y - b\mathcal{X} - a\mathcal{Z}, \mathcal{I}) = 0 \implies b = \beta. \tag{7}$$

We refer to $\mathcal{Z}$ as a *nuisance regressor*. If we use both a nuisance regressor $\mathcal{Z}$ and condition on a variable $\mathcal{B}$, the conditions become

$$\text{there exists } \alpha \text{ s.t. } \mathbb{E}[\mathrm{cov}(Y-\beta\mathcal{X}-\alpha\mathcal{Z}, \mathcal{I}|\mathcal{B})] = 0 \quad \text{and} \quad \mathbb{E}[\mathrm{cov}(Y-b\mathcal{X}-a\mathcal{Z}, \mathcal{I}|\mathcal{B})] = 0 \implies b = \beta,$$

and we write $\mathrm{NIV}_{\mathcal{X}\to Y}(\mathcal{I}, \mathcal{Z}|\mathcal{B})$; this corresponds to extracting the entries relevant for $\mathcal{X}$ from the output of $\mathrm{CIV}_{\mathcal{X}\cup\mathcal{Z}\to Y}(\mathcal{I}|\mathcal{B})$; by choosing $\mathcal{Z} = \emptyset$, NIV extends CIV. When $\mathcal{Z} \neq \emptyset$, we use the term *proper NIV*. The following theorem proves that requirements (CIV1) to (CIV3) are sufficient to establish identifiability of NIV.

Theorem 2 (Nuisance IV)**.** *Consider a linear SCM (see Appendix F.1.3) over variables V, and let $\mathcal{I}$, $\mathcal{X}$, $\mathcal{Z}$, $\mathcal{B}$, $\{Y\} \subseteq V$ be disjoint collections of variables from V, and let $\mathcal{G}$ be the corresponding DAG. Assume that $\mathcal{I}, \mathcal{X}, \mathcal{Z}$ and Y have zero mean and finite second moment and let β and α be the causal coefficients with which $\mathcal{X}$ and $\mathcal{Z}$ enter the structural equation for Y, respectively (some of the entries of β and α can be zero, so not all variables in $\mathcal{X}$ and $\mathcal{Z}$ have to be parents of Y). Let $\tilde{\mathcal{X}} := \mathcal{X} \cup \mathcal{Z}$. If requirements (CIV1) to (CIV3) are satisfied in $\mathcal{G}$ for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$ and Y, the causal effect β of $\mathcal{X}$ on Y is identified by* $\mathrm{NIV}_{\mathcal{X}\to Y}(\mathcal{I}, \mathcal{Z}|\mathcal{B})$.

Even though this is a straight-forward extension of IV regression, we are not aware of any work describing the idea of NIV. It will prove useful in the time series setting and even in the i.i.d. setting it is a strict generalization of CIV: there are graphs, such as the one in Fig. 3 (middle), where the causal effect $X \to Y$ is neither identified by IV nor by CIV.

For some graphs the effect $X \to Y$ can be identified by (proper) NIV and (proper) CIV. For example, in the graph in Fig. 3 (right) the effect $X \to Y$ can be identified both by $\mathrm{CIV}_{X\to Y}(I|B)$ and by $\mathrm{NIV}_{X\to Y}(\{I, B\}, Z)$. When estimated from a finite sample, the two resulting estimators are not identical, and, as the following proposition establishes, the two approaches cannot in general be sorted in terms of asymptotic variance.[14]

Proposition 1. *If an effect can be identified by CIV and by NIV, then the estimators cannot be strictly sorted in terms of asymptotic variance. More specifically, there exist data generating processes, for which CIV has strictly smaller asymptotic variance and others, for which NIV has strictly smaller asymptotic variance.*

The idea of NIV can be naturally applied in time series settings, too. In Section 4.2 we show that causal effects in VAR(1) processes as described in Section 2 can be estimated both by CIV and NIV. To establish this result, we first develop a marginalization technique of time series graphs, which allows us to apply the above results.

4. Instrumental Time Series Regression

4.1. Time series reduction

In Section 2.2, a VAR process is represented by its full time graph (see, e.g, Fig. 1). We now show that instruments and conditioning sets for VAR processes can be found by considering marginalized time graphs, which are obtained by marginalization of the full time graph to a finite set of nodes; they resemble latent projections [e.g., Verma, 1991] but are projections of graphs that are not finite.

Definition 1. Consider a process $S = [S_t]_{t\in\mathbb{Z}}$ satisfying Assumption (A1) and let $\mathcal{G}_{\text{full}}$ be the full time graph of S as defined in Section 2.2. Let $M = \{S^{i_1}_{t_1}, \dots, S^{i_m}_{t_m}\}$ be a finite collection of nodes in $\mathcal{G}_{\text{full}}$. The *marginalized time graph*, $\mathcal{G}_M$, is the graph over nodes M where for all $i, j \in M$ there is:

1. a directed edge $i \to j$ if and only if $i \to j$ in $\mathcal{G}_{\text{full}}$ or there exists $m_1 \in \mathbb{N}$, $v_1, \dots, v_{m_1} \notin M$ and a directed path $i \to v_1 \to \cdots \to v_{m_1} \to j$ in $\mathcal{G}_{\text{full}}$, and

2. a bidirected edge $i \leftrightarrow j$ if and only if there exists $m_1, m_2 \in \mathbb{N}$, $v_1, \dots, v_{m_1}, w_1, \dots, w_{m_2}, U \notin M$ in $\mathcal{G}_{\text{full}}$ such that there exists directed paths $U \to v_1 \to \cdots v_{m_1} \to i$ and $U \to w_1 \to \cdots w_{m_2} \to j$.

[14] In the i.i.d. setting, the asymptotic variances of both NIV and CIV estimators can be described by closed form expressions, see Appendix F.2.1.

The following theorem establishes that the CIV conditions being satisfied in a marginalized time graph implies a moment condition that can be used for identifying the causal effect. In the previous section, we stated identifiability results in terms of the causal effect β. The following theorem is stated in terms of the total causal effect (see Section 2.1); this generalizes the causal effect and may for instance be interesting if some predictors X are unobserved or are observed but cannot be intervened on. To do so, we state a slight modification of requirement (CIV1).

(CIV1') $\mathcal{I}$ and Y are d-separated given $\mathcal{B}$ in the graph where we remove an edge outgoing from $\mathcal{X}$ if the edge lies on a directed path from $\mathcal{X}$ to Y.

Theorem 3 (Time series IV by marginalization)**.** *Consider a process $S = [S_t]_{t\in\mathbb{Z}}$ satisfying Assumption (A1) with full time graph $\mathcal{G}_{full}$. Let Y be some node in $\mathcal{G}_{full}$ and let $\mathcal{X}, \mathcal{I}, \mathcal{Z}$, and $\mathcal{B}$ be disjoint collections of nodes from $\mathcal{G}_{full}$. Let $\tilde{\mathcal{X}} := \mathcal{X} \cup \mathcal{Z}$ and define $M := \{Y\} \cup \mathcal{X} \cup \mathcal{I} \cup \mathcal{Z} \cup \mathcal{B}$. Assume that requirements (CIV1') and (CIV2) are satisfied for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$ and Y in $\mathcal{G}_M$ (see Definition 1). Then, the following three statements hold. (i) The total causal effect $[\beta, \alpha]$ of $[\mathcal{X}^\top, \mathcal{Z}^\top]^\top$ on Y satisfies the NIV moment equation*

$$\mathbb{E}[\mathrm{cov}(Y - b\mathcal{X} - a\mathcal{Z}, \mathcal{I}|\mathcal{B})] = 0. \tag{8}$$

(ii) Further, if requirement (CIV3) is satisfied for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$, then $[\beta, \alpha]$ is the unique solution to (8). (iii) If, additionally, $\mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{Z}$ and $\mathbf{B}$ are observations of $\mathcal{X}, Y, \mathcal{I}, \mathcal{Z}$ and $\mathcal{B}$ at T time points, W is a positive definite matrix, and

$$[\hat{b}, \hat{a}] := \arg\min_{b,a} \|\hat{\mathrm{cov}}(\mathbf{Y} - b\mathbf{X} - a\mathbf{Z}, \mathbf{I}|\mathbf{B})\|_W^2, \tag{9}$$

then $\hat{b}$ is a consistent estimator for β.

We now apply the above result to VAR(1) processes satisfying Assumption (A2). In this case, the total causal effect coincides with the β defined in Section 2.1 (and (CIV1') and (CIV1) become equivalent). The generality of Theorem 3, however, can be used to develop similar results for VAR(p) processes with $p \neq 1$ and for total causal effects between arbitrary variables in the process.

4.2. Instrumental first-order VAR processes

We now consider estimating the causal effect β in VAR(1)-processes, such as the one displayed in Fig. 4. A first attempt might be to estimate the effect β from X_{t-1} to Y_t by directly adapting the i.i.d. case and using $\mathrm{IV}_{X_{t-1}\to Y_t}(I_{t-2})$. In Proposition 2, we prove that this estimator, in general, is not consistent. Instead, we will later make use of the time series reduction introduced above, which motivates two different (and consistent) estimators.

Proposition 2 (Failure of naive IV adaption)**.** *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying Assumption (A2) with $d_I = d_X = d_H = d_Y = 1$. If $\mathrm{cov}(X_{t-1}, I_{t-2}) \neq 0$ and*

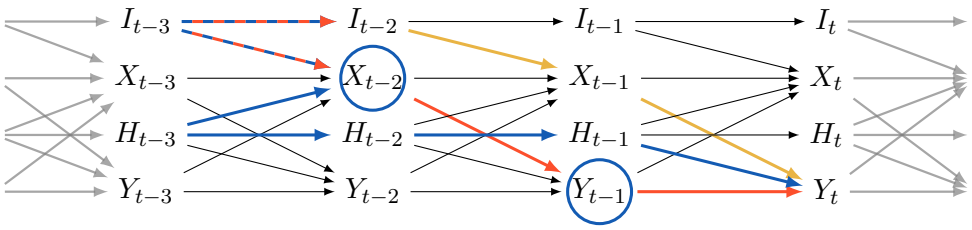


Figure 4.: Part of the full time graph of a process satisfying Assumption (A2) and three different paths from I_{t-2} to Y_t. To use I_{t-2} as an instrument to identify the causal effect β of X_{t-1} on Y_t, it is required that the only open path between I_{t-2} and Y_t be the yellow path, $I_{t-2} \to X_{t-1} \to Y_t$. If we do not use any blocking set, also the red path is unblocked. Adding X_{t-2} and Y_{t-1} as a conditioning set (blue circles) does not suffice, since the blue path is unblocked given X_{t-2} and Y_{t-1}, for the reason that X_{t-2} acts as a collider [Pearl, 2009]. By using time series reduction, we construct consistent estimators using conditional and nuisance IV (Sections 4.2.1 and 4.2.2).

$\alpha_{I,I}\alpha_{Y,Y} \neq 1$, *the* $\mathrm{IV}_{X_{t-1}\to Y_t}(I_{t-2})$ *estimator* $\hat{\beta}$ *converges in probability to*

$$(1 - \alpha_{I,I}\alpha_{Y,Y})^{-1}\beta.$$

Consequently, $\hat{\beta}$ *is in general not consistent for the causal effect* β *of* X_{t-1} *on* Y_t*, unless* I *or* Y *do not have any autoregressive structure, that is,* $\alpha_{I,I} = 0$ *or* $\alpha_{Y,Y} = 0$.

This naive IV approach fails due to the memory of the processes: The instrument, I_{t-2}, correlates with the response, Y_t, not only through the directed path $I_{t-2} \to X_{t-1} \to Y_t$ (yellow path in Fig. 4), but also through paths reaching into the past, such as $I_{t-2} \leftarrow I_{t-3} \to X_{t-2} \to Y_{t-1} \to Y_t$ (red path in Fig. 4). While additionally including the variables X_{t-2} and Y_{t-1} (circled blue) as conditioning sets would block the red path it also unblocks the blue path (since marginally independent variables, such as I_{t-3} and H_{t-3}, may become dependent when conditioning on a common descendant such as X_{t-2}). Consequently, the estimation would still be inconsistent. Instead, the concepts of NIV and time series reductions introduced above provide us with a principled approach to selecting which variables to include into the regression and allow us to construct estimators that adjust for the past.

4.2.1. Blocking the past using conditional IV

Consider a process S satisfying Assumption (A2). By Theorem 3, any set satisfying requirements (CIV1) to (CIV3) in the corresponding marginalised time graph yields a consistent estimator for the causal effect of X_{t-1} on Y_t. Thus, define $M := \{I_{t-3}, I_{t-2}, X_{t-2}, X_{t-1}, Y_{t-1}, Y_t\}$ and consider the marginalization $\mathcal{G}_M$ of the full time graph with respect to M, see Fig. 5 (left). In this graph, every path from I_{t-2} to Y_t either goes through X_{t-1} or I_{t-3}. Indeed, the assumptions for Theorem 3 are satisfied when choosing $\mathcal{I}_t := \{I_{t-2}\}$ and either

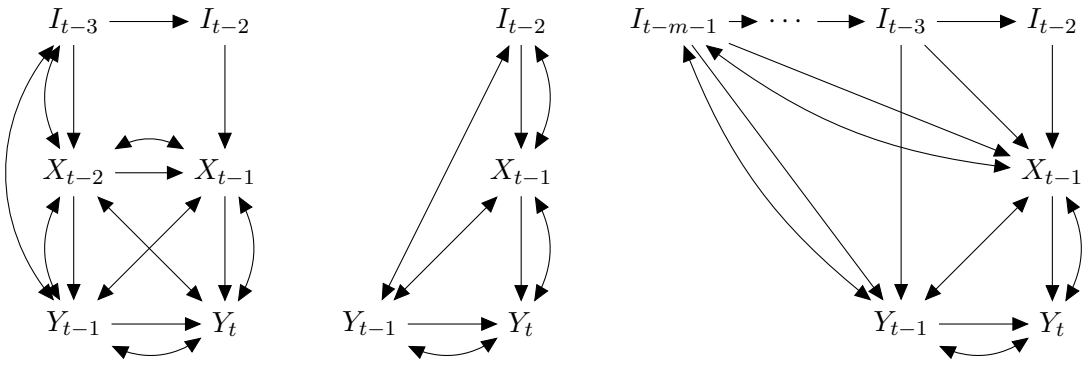


Figure 5.: Three different marginalizations of the full time graph $\mathcal{G}_{\text{full}}$ of a process satisfying Assumption (A2) (see Fig. 1). Using Theorem 3, we use these marginalizations for identification using CIV (left) and NIV (middle and right). (*Left*) Marginalization to nodes I_{t-2}, X_{t-1} and Y_t and their lagged values. (*Middle*) Marginalization to nodes I_{t-2}, X_{t-1} and Y_t and the lagged value of Y_t. (*Right*) Marginalization to m instrument nodes $I_{t-2}, \dots, I_{t-m-1}$, and X_{t-1}, Y_t, and Y_{t-1}.

$\mathcal{B}_t = \{I_{t-3}\}$ or $\mathcal{B}_t := \{I_{t-3}, X_{t-2}, Y_{t-1}\}$. Formally, we have the following theorem.

Theorem 4 (Identification with conditioning set)**.** *Consider a process* $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ *satisfying Assumption (A2). Let either* $\mathcal{B}_t := \{I_{t-3}\}$ *or* $\mathcal{B}_t := \{I_{t-3}, X_{t-2}, Y_{t-1}\}$*. Then, the following three statements hold. (i) The causal effect* β *of* X_{t-1} *on* Y_t *satisfies the CIV moment condition* $\mathbb{E}[\text{cov}(Y_t - \beta X_{t-1}, I_{t-2}|\mathcal{B}_t)] = 0$*. (ii) Furthermore, if* $\mathbb{E}[\text{cov}(X_{t-1}, I_{t-2}|\mathcal{B}_t)]$ *has rank* d_X*, then* β *is identified by* $\text{CIV}_{X_{t-1}\to Y_t}(I_{t-2}|\mathcal{B}_t)$*. (iii) If, additionally,* $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{I}_t$*, and* $\mathbf{B}_t$ *are observations of* X, Y, I *and* $\mathcal{B}$ *at* T *time points, then* β *can be consistently estimated as* $T \to \infty$ *by* $\text{CIV}_{\mathbf{X}_{t-1}\to\mathbf{Y}_t}(\mathbf{I}_{t-2}|\mathbf{B}_t)$*, that is, the output of Algorithm 1.*

Theorem 4 establishes that the bias due to confounding from the past (see beginning of Section 4.2) can be overcome by choosing either $\mathcal{B}_t = \{I_{t-3}\}$ or $\mathcal{B}_t = \{I_{t-3}, X_{t-2}, Y_{t-1}\}$ as conditioning set. In Section 5, we compare these two choices empirically. The assumption that α_{XI} has full rank ensures the relevance condition, requirement (CIV3). Other choices of the marginalization set M are possible, too, and yield alternative ways of estimating the causal effect $X_{t-1} \to Y_t$: We now illustrate an alternative strategy for identification using nuisance IV.

4.2.2. Blocking the past using nuisance IV

The graph in Fig. 5 (*middle*) shows the marginalization of the full time graph in Fig. 4 to nodes $M := \{I_{t-2}, X_{t-1}, Y_{t-1}, Y_t\}$. The effect $X_{t-1} \to Y_t$ cannot be consistently estimated using only these variables in a CIV; if we condition on Y_{t-1}, for example, the path $I_{t-2} \leftrightarrow Y_{t-1} \leftrightarrow Y_t$ is unblocked, violating requirement (CIV1). Instead, we can include Y_{t-1} as a nuisance regressor: We identify the effect of X_{t-1} on Y_t using the instrument $\mathcal{I}'_t := \{I_{t-2}\}$ by $\text{NIV}_{X_{t-1}\to Y_t}(\mathcal{I}'_t, Y_{t-1})$, defined in Section 3.2. This model

Algorithm 1 Estimating the causal effect β of X_{t-1} on Y_t under Assumption (A2).

Input: Sample $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_T] \in \mathbb{R}^{d_X \times T}$, $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_T] \in \mathbb{R}^{d_Y \times T}$, $\mathbf{I} = [\mathbf{I}_1, \dots, \mathbf{I}_T] \in \mathbb{R}^{d_I \times T}$

1: Align observations, by setting $\mathbf{Y} = [\mathbf{Y}_s, \dots, \mathbf{Y}_T]$ and either

(CIV) $\mathbf{I} \coloneqq [\mathbf{I}_{s-2}, \dots, \mathbf{I}_{T-2}]$ $\mathbf{X} \coloneqq [\mathbf{X}_{s-1}, \dots, \mathbf{X}_{T-1}]$ and $\mathbf{B} \coloneqq [\mathbf{I}_{s-3}, \dots, \mathbf{I}_{T-3}]$

(NIV) $\mathbf{I} \coloneqq \begin{bmatrix} \mathbf{I}_{s-2} & & \mathbf{I}_{T-2} \\ \vdots & \dots & \vdots \\ \mathbf{I}_{s-m+1} & & \mathbf{I}_{T-m+1} \end{bmatrix}$ and $\mathbf{X} \coloneqq \begin{bmatrix} \mathbf{X}_{s-1} & \dots & \mathbf{X}_{T-1} \\ \mathbf{Y}_{s-1} & \dots & \mathbf{Y}_{T-1} \end{bmatrix}$,

where s is chosen such that all indices are positive.

2: Compute regression estimates $\hat{E}[\mathbf{X}|\mathbf{B}], \hat{E}[\mathbf{Y}|\mathbf{B}]$, and $\hat{E}[\mathbf{I}|\mathbf{B}]$.

3: Compute residual processes $r_{\mathbf{X}} \coloneqq \mathbf{X} - \hat{E}[\mathbf{X}|\mathbf{B}], r_{\mathbf{Y}} \coloneqq \mathbf{Y} - \hat{E}[\mathbf{Y}|\mathbf{B}]$, and $r_{\mathbf{I}} \coloneqq \mathbf{I} - \hat{E}[\mathbf{I}|\mathbf{B}]$.

4: $W \coloneqq \left(\frac{1}{T-s+1} r_{\mathbf{I}} r_{\mathbf{I}}^\top\right)^{-1}$

5: $\hat{\beta} \coloneqq r_{\mathbf{Y}} r_{\mathbf{I}}^\top W r_{\mathbf{I}} r_{\mathbf{X}} \left(r_{\mathbf{X}} r_{\mathbf{I}}^\top W r_{\mathbf{I}} r_{\mathbf{X}}^\top\right)^{-1}$

6: If $\mathbf{I}$ and $\mathbf{X}$ are chosen according to NIV, set $\hat{\beta} \coloneqq \hat{\beta}_{1:d_X}$.

Output: Estimate of causal effect $\hat{\beta}$

satisfies the (nuissance) IV requirement (CIV1), because the only open paths between I_{t-2} and Y_t includes either the edge $X_{t-1} \to Y_t$ or the edge $Y_{t-1} \to Y_t$.

Because the resulting d_X-dimensional estimate is extracted from the $d_X + d_Y$-dimensional solution $\mathrm{IV}_{\{X_{t-1}, Y_{t-1}\} \to Y_t}(\mathcal{I}'_t)$, we require that $\mathsf{rank}\mathbb{E}\left\{[X_{t-1}^\top, Y_{t-1}^\top]^\top (\mathcal{I}'_t)^\top\right\} = d_X + d_Y$. If $d_I < d_X + d_Y$, this rank condition is not met for $\mathcal{I}'_t \coloneqq \{I_{t-2}\}$. To overcome this, one can increase the instrument set to $\mathcal{I}_t = \{I_{t-2}, I_{t-3}, \dots, I_{t-m-1}\}$: Fig. 5 (*right*) shows the marginalization of the full time graph to $M \coloneqq \mathcal{I}_t \cup \{X_{t-1}, Y_{t-1}, Y_t\}$. Again, requirements (CIV1) and (CIV2) are satisfied in $\mathcal{G}_M$, but the instrument set now has dimension $|\mathcal{I}| = md_I$ and, provided the rank condition now holds, β is identified by $\mathrm{NIV}_{X_{t-1} \to Y_t}(\mathcal{I}_t, Y_{t-1})$. The following theorem formalizes this discussion.

Theorem 5 (Identification with nuisance regressor)**.** *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t \in \mathbb{Z}}^\top$ satisfying Assumption (A2). Let $\mathcal{I}_t \coloneqq \{I_{t-2}, \dots, I_{t-m-1}\}$ for an $m \geq 1$ and $\mathcal{Z}_t \coloneqq \{Y_{t-1}\}$. Then, the following three statements hold. (i) There exists $\alpha \in \mathbb{R}$ such that the causal effect β of X_{t-1} on Y_t satisfies the NIV moment condition $\mathbb{E}[\mathrm{cov}(Y_t - \beta X_{t-1} - \alpha \mathcal{Z}_t, \mathcal{I}_t)] = 0$. (ii) Further, if $\mathbb{E}[[X_{t-1}^\top, \mathcal{Z}_t^\top]^\top \mathcal{I}_t^\top]$ has rank $d_X + d_Y$, β is identified by $\mathrm{NIV}_{X_{t-1} \to Y_t}(\mathcal{I}_t, \mathcal{Z}_t)$. (iii) If, additionally, $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{I}_t$, and $\mathbf{Z}_t$ are observations of $X, Y, \mathcal{I}$ and $\mathcal{Z}$ at T time points, then β can be consistently estimated as $T \to \infty$ by $\mathrm{NIV}_{\mathbf{X}_{t-1} \to \mathbf{Y}_t}(\mathbf{I}_t, \mathbf{Z}_t)$, that is, the output of Algorithm 1.*

Theorem 5 shows that identification is possible if $\mathsf{rank}\mathbb{E}\left\{[X_{t-1}^\top, Y_{t-1}^\top]^\top \mathcal{I}_t^\top\right\} = d_X + d_Y$

is met, where we use the lagged instrument set $\mathcal{I}_t = \{I_{t-2}, \dots, I_{t-m-1}\}$ (it is easy to see that for this choice of $\mathcal{I}_t$, identifiability also holds for $\beta = 0$ and the rank being d_X). Satisfying this relevance criterion implies requirement (CIV3). For the CIV approach in Section 4.2.1, this is directly related to the rank of the parameter α_{XI}. For NIV, the relevance criterion depends on the parameter matrix in an intricate way. We now provide necessary and sufficient conditions for when this rank condition is satisfied. Moreover, we show in Corollary 1 below that for almost all parameter matrices one can obtain sufficiently high rank to identify the effect $X_{t-1} \to Y_t$ by increasing the number of lags used.

Define the following submatrices of the parameter matrix A from Assumption (A2).

$$A_I := \begin{pmatrix} \alpha_{X,I} \\ 0 \end{pmatrix} \in \mathbb{R}^{d_X+1} \quad \text{and} \quad A_{XY} := \begin{pmatrix} \alpha_{X,X} & \alpha_{X,Y} \\ \beta & \alpha_{Y,Y} \end{pmatrix} \in \mathbb{R}^{(d_X+1)\times(d_X+1)}. \tag{10}$$

The following theorem outlines conditions for $\mathbb{E}[[X_{t-1}^\top, Y_{t-1}^\top]^\top \mathcal{I}^\top]$ to have full rank when $d_I = 1$ and we use $d_X + d_Y = d_X + 1$ lags as instruments:

Theorem 6. *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying Assumption (A2). Assume that $d_I = d_Y = 1$ and let $\mathcal{I}_t := \{I_{t-2}, \dots, I_{t-m-1}\}$, where $m = d_X + d_Y$. Let A_{XY} and A_I be defined as in (10). The following three statements are equivalent:*

1. $\mathsf{rank}\mathbb{E}[[X_{t-1}^\top, Y_{t-1}^\top]^\top \mathcal{I}_t^\top] = d_X + d_Y$.
2. *The matrix $\left[A_{XY}^0 A_I,\, A_{XY}^1 A_I, \dots, A_{XY}^{d_X} A_I\right]$ is invertible, where A_{XY}^0 is the identity matrix of size $(d_X + d_Y) \times (d_X + d_Y)$.*
3. *Different Jordan blocks of J have different eigenvalues and for all $q \in \{1, \dots, k\}$, the coefficient $w_{\sum_{i=1}^q m_i}$ is non-zero; here, $J = Q^{-1} A_{XY} Q$ is the Jordan normal form[15] of A_{XY}, with k Jordan blocks $J = \mathrm{diag}(J_{m_1}(\lambda_1), \dots, J_{m_k}(\lambda_k))$, each with size m_i and eigenvalue λ_i, and w are the coefficients of A_I in the basis of the generalized eigenvectors Q, that is, $w = Q^{-1} A_I$.*

Intuitively, Theorem 6 states that identification of β is possible if the information is passed on in a sufficiently diverse way from I_{t-k} to (X_t, Y_t). For every $k \in \{0, \dots, d_X\}$, $A_{XY}^k A_I$ corresponds to the path $I_{t-(k+1)} \overset{A_I}{\to} (X_{t-k}, Y_{t-k}) \overset{A_{XY}}{\to} \dots \overset{A_{XY}}{\to} (X_t, Y_t)$ in Fig. 1. Theorem 6 2 requires these to be sufficiently different (that is, linearly independent), for the matrix $\mathbb{E}[[X_{t-1}^\top, Y_{t-1}^\top]^\top \mathcal{I}_t^\top]$ to have full rank. While there are parameter matrices that do not satisfy Theorem 6 (see Appendix F.3.1 for examples), the following corollary shows that, when chosen randomly, almost all parameter matrices A allow for using multiple lags I_{t-2-j} as instruments.

Corollary 1. *Consider a VAR(1) process S with $d_I = 1$ and parameter matrix A, and assume that sparsity pattern of A is given by Assumption (A2) and that the non-zero entries of A are drawn from any distribution which has density with respect to Lebesgue measure. Then β is identifiable with probability* 1.

[15] See Appendix F.4.9 for the definition of Jordan normal forms and the notation that we use.

The following corollary provides a sufficient condition for identifiability of β, when instruments are multivariate.

Corollary 2. *Consider a process S satisfying Assumption (A2) with $d_I > 1$ instrument processes $I^{(1)}, \dots, I^{(d_I)}$. Assume that there is at least one instrument process $I^{(j)}$ such that both of the following conditions hold.*

1. *$I_t^{(j)}$ is independent of $I_s^{(i)}$ for all t, s and $i \neq j$, and*
2. *the requirements of Theorem 6 are satisfied for the reduced process $(I^{(j)}, X, Y)$.*

Then β is identifiable.

Using a single instrument at $d_X + 1$ lags allows for a simple condition for identifiability. But in finite samples, using instruments with high time lags may come at a loss of efficiency as the estimation procedure may suffer from weak dependencies between the residual and the instrument due to the mixing of the time series, see Section 5 for an empirical investigation.

4.2.3. Extension to non-VAR Processes

The results in Sections 4.2.1 and 4.2.2 assume that the entire process S is a VAR(1) process, see Assumption (A2), which was done mostly for presentation purposes. The key arguments and statements hold more generally. We have argued that similar arguments hold for VAR(p) processes and we now consider a setting, where Y_t satisfies a linear structural equation (as in a VAR(1) process) but we do not assume any VAR structure or linearity on the remaining subprocesses. We outline the assumptions needed to obtain the same identification results as in Sections 4.2.1 and 4.2.2 (a similar relaxation can be applied when Y_t behaves like a VAR(p) process). Let $[S_t]_{t\in\mathbb{Z}} = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ and assume that for all $t \in \mathbb{Z}$

$$Y_t := \beta X_{t-1} + \alpha_{Y,Y} Y_{t-1} + g(\varepsilon_t^Y, H_{t-1}), \tag{11}$$

where ε_t^Y is a sequence of i.i.d. random variables that, for all t, are independent of $[Y_s]_{s<t}$ and $[X_s, I_s, H_s]_{s\leq t}$ and g is a measurable function. Without assuming a VAR process we make the following assumptions on S.

(A1') $[S_t]_{t\in\mathbb{Z}}$ is covariance stationary.

(A2') S satisfies (2), that is, empirical first and second moments converge to their population version.

(A3') There exists a $p \in \mathbb{N}$ such that for all $i \in \{1, \dots, d\}$, there exists a function f^i such that for all $t \in \mathbb{Z}$

$$S_t^i = f^i(S_{t-1}, ..., S_{t-p}) + \varepsilon_t^i.$$

This induces a full time graph $\mathcal{G}_{\text{full}}$ [Peters et al., 2013]; we assume that for all finite disjoint collections of nodes A, B, C from $\mathcal{G}_{\text{full}}$ such that A and C are d-separated given B in $\mathcal{G}_{\text{full}}$, we have $A \perp\!\!\!\perp C|B$. Furthermore, for all $t \in \mathbb{Z}$, $H_{t-1} \in \text{ND}(I_{t-2})$, $Y_{t-1} \in \text{ND}(I_{t-2})$ and ε_Y^t is independent of any finite set $A \subseteq \text{ND}(Y_t)$.

(A4') For all $t \in \mathbb{Z}$ and $m \in \mathbb{N}$, we have $(\varepsilon_t^Y, H_{t-1}) \perp\!\!\!\perp (I_{t-2}, ..., I_{t-2-m})$.

Using these assumptions, we can restate Theorem 4 without the VAR(1) assumption.

Proposition 3 (Identification with conditioning set relaxing the VAR assumption)**.** *Consider a process* $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ *satisfying* (11) *and Assumptions (A1') to (A3'). Let* $\mathcal{B}_t$ *be a set of variables satisfying* $\text{PA}(I_{t-2}) \subseteq \mathcal{B}_t \subseteq \text{ND}(Y_t) \cap \text{ND}(I_{t-2})$ *in* $\mathcal{G}_{full}$*. Then, (i), (ii) and (iii) from Theorem 4 hold.*

Similarly, we can extend Theorem 5 to more general time series models, too.

Proposition 4 (Identification with nuisance regressor relaxing the VAR assumption)**.** *Consider a process* $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ *satisfying* (11) *and Assumptions (A1'), (A2') and (A4'). Let* $\mathcal{Z}_t := \{Y_{t-1}\}$ *and* $\mathcal{I}_t := \{I_{t-2}, \dots, I_{t-m-1}\}$ *for an* $m \geq 1$*. Then, (i), (ii), and (iii) from Theorem 5 hold.*

So far, we have focused on estimating causal effects. Knowledge of such causal effects can be of interest in itself. In the following section, we discuss that they can also be used for prediction and forecasting under interventions.

4.3. Optimal prediction under interventions

Causal estimates may facilitate improved prediction under intervention. Suppose that we have inferred the causal effect $X_{t-1} \to Y_t$ for instance using the methods presented in Section 4.2 above. How do we best predict Y_{t+1} given that we perform the intervention $\text{do}(X_t := x)$ (see Appendix F.1.3 for an introduction to do-interventions) and that we observe the past values of the time series (but the value of X_t had it not been intervened on is not observed)?

Due to the hidden confounding, the conditional mean of Y_{t+1} given its past (which could be consistently estimated by OLS regression, for example) is in general not optimal in terms of mean square prediction error (MSPE). Intuitively, this is because the conditional mean also encompasses the effects of the latent process H onto Y_{t+1}. In the i.i.d. setting, it has been observed that using the causal parameter yields a smaller MSPE and can be worst-case optimal under arbitrarily large interventions [e.g., Rojas-Carulla et al., 2018, Christiansen et al., 2021].

In VAR processes, the intervention $\text{do}(X_t := x)$ partially breaks the confounding of X_t and Y_{t+1} from the past, yet, due to the latent process H, the process (X, Y) is not a Markov process such that the lagged observations $\{X_{t-k}, k = 1, \dots, m\}$ and $\{Y_{t-j}, j = 1, \dots, l\}$ further improve prediction of Y_{t+1}. Fixing the number of lags, the following proposition shows that, under the intervention $\text{do}(X_t := x)$, the optimal linear prediction consist of a mix of (population) regression parameters for non-intervened variables and causal parameters for the intervened variable X_t.

Proposition 5. *Consider a process $S = [S_t]_{t\in\mathbb{Z}}$ satisfying Assumption (A2). Let β be the causal effect from X_t to Y_{t+1}, and let for an arbitrary $m, \ell \in \mathbb{N}$ $(\alpha_{Y,X}, \alpha_{Y,Y})$ be the population vector of coefficients when regressing $Y_{s+1} - \beta X_s$ on $\{X_{s-k}, k = 1, \dots, m\} \cup \{Y_{s-j}, j = 0, \dots, l\}$. Then*

$$(\alpha_{Y,Y}, \beta, \alpha_{Y,X}) = \arg\min_{a,b,c} \mathbb{E}_{\mathrm{do}(X_t:=x)} \left\{ Y_{t+1} - \sum_{j=0}^{\ell} a_j Y_{t-j} - bX_t - \sum_{k=1}^{m} c_k X_{t-k} \right\}^2 .$$

That is, under the intervention $\mathrm{do}(X_t := x)$, the causal coefficient can be used to optimally predict Y_{t+1}. We state the corresponding finite sample algorithm in Appendix F.3.2.

5. Simulation Experiments

We test the empirical performance of our proposed estimators in simulation experiments.

Data generating process. We generate data by first simulating a matrix A with the sparsity structure of Fig. 2a and all non-zero entries being drawn independently uniformly at random from $(-0.9, -0.1) \cup (0.1, 0.9)$. By Corollary 1, the causal effect under this sampling scheme is almost surely identifiable by the NIV. Unless specified differently, we use $d_I = 3, d_X = 2$ and $d_Y = 1$. Any such randomly generated matrix is used to generate data only if it satisfies the eigenvalue condition in Assumption (A1) with a margin of 0.1. Furthermore, all noise variables ε_t^i are independently randomly drawn from a normal distribution with mean 0 and standard deviation 1.

Evaluation of the estimators. We simulate $m = 1{,}000$ random matrices and for each we simulate $s = 10$ data sets. We fit estimators $\hat{\beta}_j$ for each $j \in \{1, \dots, s\}$, and for a given matrix, we compute the average error, $\mathrm{error}(\hat{\beta}) := \mathrm{mean}(\|\hat{\beta}_1 - \beta\|_2^2, \dots, \|\hat{\beta}_s - \beta\|_2^2)$.

5.1. Identification of the causal effect

Consistency The estimators CIV and NIV are consistent (Theorems 4 and 5) and we perform an experiment to compare their finite sample properties for different sample sizes. We simulate data from the scheme described above and fit two $\mathrm{CIV}_{X_{t-1}\to Y_t}(I_{t-2}|\mathcal{B}_t)$ estimators, where $\mathcal{B}_t = \{I_{t-3}\}$ (CIV_I) or $\mathcal{B}_t = \{I_{t-3}, X_{t-2}, Y_{t-1}\}$ ($\mathrm{CIV}_{I,X,Y}$) and two $\mathrm{NIV}_{X_{t-1}\to Y_t}(\mathcal{I}_t, Y_{t-1})$, where $\mathcal{I}_t = \{I_{t-2}\}$ ($\mathrm{NIV}_{1\text{ lag}}$) or $\mathcal{I}_t = \{I_{t-2}, I_{t-3}, I_{t-4}\}$ ($\mathrm{NIV}_{3\text{ lag}}$). All of these estimators are consistent (see Section 4.2). We plot the errors obtained for different sample sizes T in Fig. 6 (left). For all estimators, the errors decrease with increasing sample size supporting the consistency results in Theorem 3. In general, there is no empirical support for either of the NIV or CIV estimators to be strictly better than the others in terms of speed of convergence in terms of sample size. As discussed in Section 4.2.1, both $\mathcal{B}_t = \{I_{t-3}\}$ and $\mathcal{B}_t = \{I_{t-3}, X_{t-2}, Y_{t-1}\}$ block confounding from past values. We observe that removing X_{t-2} and Y_{t-1} from $\mathcal{B}_t$ increases the upper tail

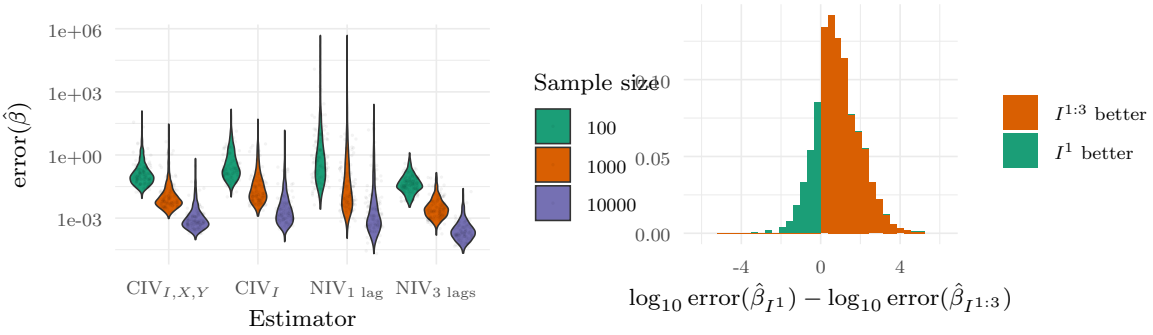


Figure 6.: (*Left*) Distributions of the average error (in log scale) of different consistent CIV and NIV estimators for various sample sizes T, see Section 5.1 ('Consistency'). Each point corresponds to the average over repeated draws from the same parameter matrix, and different points correspond to different parameter matrices. The $\text{NIV}_{1\text{ lag}}$ estimators have heavier tails, corresponding to the fact that this is a just-identified case, where we use a 3-dimensional instrument to estimate a 3-dimensional causal effect. (*Right*) Histogram of log error ratios for two different NIV estimators in a model with a 3-dimensional instrument. I^1 uses 6 lags from a 1-dimensional instrument process, while $I^{1:3}$ uses 2 lags of a 3-dimensional instrument process, see Section 5.1 ('Using more lags or additional instruments'). A value larger than zero indicates that I^1 yields a larger error than $I^{1:3}$. This is the case for the majority of the considered settings.

of the error distribution, supporting the intuition that while the conditioning variables X_{t-2} and Y_{t-1} are not necessary for identification, they reduce finite sample variance. Similarly, for the NIV estimators, using 3 lags instead of a single lag shrinks the upper tail of the error distribution.

Using more lags or additional instruments. We compare using multiple lags of a single instrument to using multiple instrument processes. We consider the model described above with $d_I = 3$ independent instrument processes, and use either the first instrument process (I^1) or all three instrument processes ($I^{1:3}$) for estimation in NIV. For $I^{1:3}$ we use 2 lags of each of the three processes ('recent instruments'), while for I^1 we use 6 lags ('distant instruments'), such that both models use in total 6 instruments. In this way, we can inspect the benefit of using more recent instruments if available.

Figure 6 (right) shows a histogram of the log error ratio of the two different estimators: A large value indicates that model I^1, which uses many lags of only a single instrument process, incurs a higher error. In the majority of parameter settings, using recent instruments yields a lower error, in some cases several orders of magnitude, when compared to the more distant instruments. Although adding lags of a univariate instrument process can yield identifiability of the causal effect of a regressor process that is not

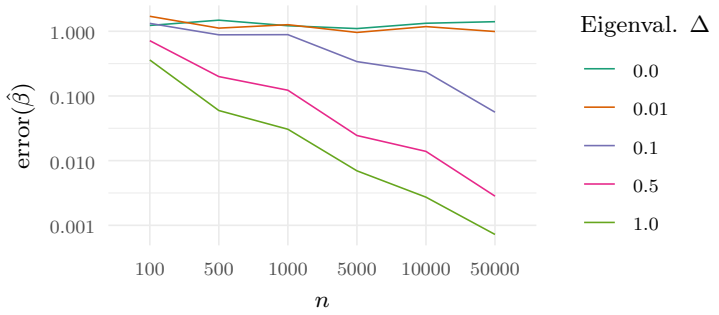


Figure 7.: Median error (log scale) for the NIV estimator as we vary Δ, the difference between the eigenvalues of $\alpha_{X,X}$, see Section 5.1 ('Estimation close to non-identifiability'). The causal effect is identifiable if and only if $\Delta \neq 0$, see Theorem 6. Indeed, the error does not decrease for $\Delta = 0$ and decreases the faster, the further Δ is away from 0.

univariate (see Theorem 6), this simulation experiment supports the notion that using more instruments (if available) is preferred over using more distant lags.

Estimation close to non-identifiability. Example F.1 in Appendix F.3.1 shows a setting, in which β is not identifiable by NIV (in the setting of Corollary 1, this happens with probability zero). We examine the behaviour of the NIV estimator in scenarios that are close to this non-identifiable setting. We consider $d_I = 1$, $\alpha_{X,X} = \text{diag}(-0.6, -0.6 + \Delta)$ and draw the remaining parameters uniformly from $(-0.9, -0.1) \cup (0.1, 0.9)$. As per Corollary 1, the causal effect is identifiable, except for $\Delta = 0$, in which case $\alpha_{X,X}$ has two Jordan blocks with the same eigenvalue. In Fig. 7 we plot the median error for changing Δ and sample size T.[16] The further Δ is from 0, the faster (in terms of sample size) the estimator converges to β. In the non-identified setting $\Delta = 0$, we do not observe the error to decrease with increasing sample size. This observation is in line with Theorem 6.

5.2. Using the causal parameter for prediction under intervention on X

We support empirically that a linear prediction using OLS parameters for non-intervened variables and causal parameters for the intervened variables achieves minimal square loss for prediction under intervention (see Proposition 5). We consider the model with $d_X = d_I = 1$ and, ensuring strong hidden confounding, draw the non-zero entries $\alpha_{i,H}$, $i \in \{X, H, Y\}$ uniformly at random from $(-0.9, -0.5) \cup (0.5, 0.9)$ (instead of $(-0.9, -0.1) \cup (0.1, 0.9)$ as for the other non-zero entries of A). The prediction task follows Section 4.3: Given observations $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{T-1}] \in \mathbb{R}^{d_X \times T-1}$ and $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_T] \in \mathbb{R}^{d_Y \times T}$, with $T = 3{,}000$, the goal is to predict $\mathbf{Y}_{T+1} \in \mathbb{R}^{d_Y \times 1}$ under an intervention $\text{do}(X_T :=$

[16] Here, we report the median, since the non-identifiability when $\Delta = 0$ implies that the mean is ill-behaved; for those lines where $\Delta \neq 0$, plotting the mean instead of the median yields a similar plot (not shown).

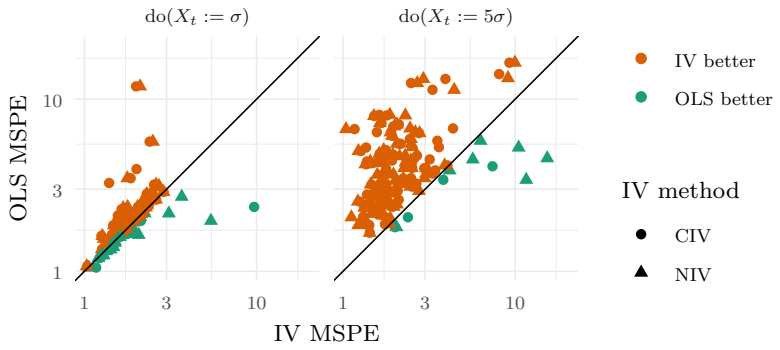


Figure 8.: The figure plots the loss MSPE($\hat{Y}_{T+1}$) under an intervention do($X_T := x$) when Y_{T+1} is predicted using OLS against the loss when Y_{T+1} is predicted using one of the IV methods we develop for time series. The OLS prediction is based on the regression $Y_{t+1} \sim X_t + X_{t-1} + X_{t-2} + Y_t + Y_{t-1}$ while the IV predictions of Y_{T+1} are based on the procedure discussed in Section 4.3 (see Algorithm F.1 in Appendix F.3.2). We plot this both for do($X_T := \sigma$) and do($X_T := 5\sigma$), where σ is the standard deviation of X_t in the unintervened distribution. The results show that as the intervention strength increases, the OLS prediction error increases at a faster rate than the IV prediction error. 17 outliers were removed from the right-hand side plot, of which 10 had a larger OLS MSPE than IV MSPE.

$n \cdot \sigma$) where σ is the standard deviation the process $[X_t]_{t=1}^{T-1}$ and $n \in \{1, 5\}$. In Section 4.3 we discuss that IV is prediction optimal under arbitrary interventions do($X_T := x$). In particular, one would expect that OLS becomes increasingly inferior to IV methods when one increases n and thereby the intervention strength do($X_T := n \cdot \sigma$).

We compare prediction for Y_{T+1} via Algorithm F.1 in Appendix F.3.2 (with $m = 2$ and $l = 1$) with prediction based on the OLS regression $Y_{t+1} \sim X_t + X_{t-1} + X_{t-2} + Y_t + Y_{t-1}$ ('OLS'). For each repetition and matrix A, we obtain CIV and NIV estimates of β on a separate sample first, and then obtain predictions for Y_{T+1} following Algorithm F.1 using either $\hat{\beta}_{\text{CIV}}$ ('CIV') or $\hat{\beta}_{\text{NIV}}$ ('NIV'). For each of the $m = 100$ random matrices A we compute error($\hat{Y}_{T+1}$) as the mean of the squared prediction error $(\hat{Y}_{T+1} - Y_{T+1})^2$ over the $s = 100$ repetitions.

In Fig. 8, we plot the error of the OLS estimate against the error of the IV estimate (either CIV or NIV). When ($X_T := \sigma$), the intervention is within the normal range of X_t, and the errors of OLS and IV estimates are similar. As we perform a stronger intervention ($X_T := 5\sigma$), the OLS error exceeds the IV error in most simulations. This robustness in prediction under intervention is in line with the result in Proposition 5.

6. Conclusion and Future Work

In this work, we have developed IV methods for time series data that allow us to identify the causal effect of a process X on a response process Y, based on an instrument process I that exhibits memory effects. Simple adaptations of ordinary IV estimators generally fail to identify the causal effect due to confounding from the past, as we show in Proposition 2. We have developed the concept of nuisance IV (NIV), see Theorem 2, a marginalization framework for time series graphs, see Theorem 1, and a global Markov property for VAR processes, see Theorem 3. Based on these principles, we propose two classes of estimation methods that properly adjust for confounding from the past: one based on choosing the correct conditioning set (CIV), see Theorem 4, and another one based on nuisance regressors (NIV), see Theorem 5. The procedures find solutions to moment conditions that, in their population version, are satisfied for the true causal parameters. Unlike in the i.i.d. case, the identifiability conditions (which are usually rank conditions) do not have a simple interpretation. Theorem 6 provides necessary and sufficient conditions on the parameters of the underlying data-generating process for the causal parameter being the unique solution to the corresponding moment equation.

The results of simulation experiments support the theoretical finding that the estimators are consistent. In general, different choices of instruments, conditioning sets and nuisance regressors allow us to consistently identify the causal effect but, as the experiments show, they may come with different finite sample behaviour. For example, for identifiability in the case of NIV, we only need the number of lags, m, used as instruments large enough such that $d_I \cdot m = d_X + d_Y$ holds but using more lags may help in finite samples in that this shrinks the upper tail of the error distribution, see Fig. 6 (left).

We have further argued that identifying the causal effect may be of interest not only for causal inference, but also for prediction of Y under the intervention $\mathrm{do}(X_t := x)$, where the minimal expected squared error can be obtained by a mix of causal parameters and regression coefficients, see Proposition 5 and Section 5.2.

For future work, it may be fruitful to develop principled techniques for deciding which estimator yields the best finite sample performance [see, e.g., Henckel, 2021, Chapter 4] and to construct confidence statements, either based on Appendix F.2.2 or other techniques [e.g., Newey and West, 1987, Shah and Peters, 2020]. Finally, as for the i.i.d. case [Imbens and Newey, 2009, Chesher, 2003, Saengkyongam et al., 2022], considering (higher order) independence, rather than vanishing covariances may yield stronger identifiability results but may come with computational and statistical challenges.

Acknowledgments

NT, RS and JP were supported by a research grant (18968) from VILLUM FONDEN. In addition, during parts of the project, JP and SW were supported by the Carlsberg Foundation.

A. Appendix to Statistical Testing under Distributional Shifts

A.1. Further comments on the framework

A.1.1. Forward and backward shifts, τ and η

In this paper, as mentioned in Section 2.1, we take the starting point that Q^* is observed, and view $P^* = \tau(Q^*)$ as a shifted version of Q^*. One could instead suppose that we started with a distribution of interest P^*, from which no sample is available, and then construct a map η such that $Q^* = \eta(P^*)$ is a distribution which can be sampled from in practice. If τ and η are invertible, the two views are mathematically equivalent, but if not, there is a subtle difference; the corresponding level guarantees take a supremum either over $Q \in \{Q' \mid \eta^{-1}(Q') \cap H_0 \neq \emptyset\}$ (η view) or over $Q \in \{Q' | \tau(Q') \in H_0\}$ (τ view). To see this, we first start with the (natural) level guarantee from the η view: $\sup_{P \in H_0} \mathbb{P}_{\eta(P)}(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha$. We then have

$$\begin{aligned} & \sup_{P \in H_0} \mathbb{P}_{\eta(P)}(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha \\ \Leftrightarrow \quad & \sup_{Q \in \eta(H_0)} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha \\ \Leftrightarrow \quad & \sup_{Q \in \{Q' \,|\, \eta^{-1}(Q') \cap H_0 \neq \emptyset\}} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \end{aligned}$$

If, alternatively, we start with the level guarantee from the τ view, we find

$$\begin{aligned} & \sup_{P \in H_0} \sup_{Q \in \tau^{-1}(P)} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha \\ \Leftrightarrow \quad & \sup_{Q \in \tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha \\ \Leftrightarrow \quad & \sup_{Q \in \{Q' \,|\, \tau(Q') \in H_0\}} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \end{aligned}$$

Comparing the last two lines yields the claim.

A.1.2. Example: Interventions in causal models

One example of a distributional shift τ is the case where τ represents an intervention in a structural causal model (SCM) over $X^1, \ldots, X^d$ [Pearl, 2009]. An SCM $\mathfrak{C}$ over $X^1, \ldots, X^d$ is a collection of structural assignments $f^1, \ldots, f^d$ and noise distributions

$Q_{N^1}, \dots, Q_{N^d}$ such that for each $j = 1, \dots, d$, we have $X^j := f^j(\mathrm{PA}^j, N^j)$. Here, the noise variables N^j are distributed according to $N^j \sim Q_{N^j}$ and are assumed to be jointly independent. The sets $\mathrm{PA}^j \subseteq \{X^1, \dots, X^d\} \backslash \{X^j\}$[1] denote the causal parents of X^j. The induced graph over $X^1, \dots, X^d$ is the graph obtained by drawing directed edges from each variable on the right-hand side of each assignment to the variables on the left-hand side; see Bongers et al. [2021] for a more formal introduction to SCMs.

Let us assume that $\mathfrak{C}$ induces a unique observational distribution Q over $X^1, \dots, X^d$ (which is the case if the graph is acyclic, for example), and assume that Q admits a joint density q with respect to a product measure. Then q satisfies the factorization property (see Lauritzen et al. [1990] or Theorem 1.4.1 in Pearl [2009]): $q(x^1, \dots, x^d) = \prod_{j=1}^d q_{X^j|\mathrm{PA}^j}(x^j|x^{\mathrm{PA}^j})$. In an SCM, an intervention on a variable X^k replaces the tuple $(f^k, \mathrm{PA}^k, Q_{N^k})$ with $(\bar{f}^k, \bar{\mathrm{PA}}^k, \bar{Q}_{N^j})$ in the structural assignment for X^k, and we denote the replacement by $\mathrm{do}(X^k := \bar{f}^k(\bar{\mathrm{PA}}^k, \bar{N}^k))$ [Pearl, 2009]. This new mechanism determines a conditional that we denote by $p^*(x^k|x^{\bar{\mathrm{PA}}^k})$. The interventional distribution is the induced distribution with the new structural assignment, and we denote this by $P := Q^{\mathrm{do}(X^k := \bar{f}(\bar{\mathrm{PA}}^k, \bar{N}^k))}$. If P admits the density p, only the conditional density of X^k changes [e.g., Haavelmo, 1944, Aldrich, 1989, Pearl, 2009, Peters et al., 2017], that is, for $j \neq k$, we have $p(x^j|x^{\mathrm{PA}^j}) = q(x^j|x^{\mathrm{PA}^j})$, for all x^j and x^{PA^j}. Assume that for the true but unknown distribution Q^* we know the conditional $q^*(x^k|x^{\mathrm{PA}^k})$ (e.g., because this was part of the design when generating the data). Due to the factorization property, the intervention $\mathrm{do}(X^k := \bar{f}^k(\bar{\mathrm{PA}}^k, \bar{N}^k))$ can then be represented as a map τ that acts on the density q:

$$\tau(q)(x^1, \dots, x^d) := \frac{p^*(x^k|x^{\bar{\mathrm{PA}}^k})}{q^*(x^k|x^{\mathrm{PA}^k})} \cdot q(x^1, \dots, x^d).$$

Defining $r(x^{\{k\} \cup \mathrm{PA}^k \cup \bar{\mathrm{PA}}^k}) := p^*(x^k|x^{\bar{\mathrm{PA}}^k})/q^*(x^k|x^{\mathrm{PA}^k})$, this takes the form of (5). As the conditional $p^*(x^k|x^{\bar{\mathrm{PA}}^k})$ is fully specified by the intervention, we therefore know the function r. Our proposed framework allows us to test statements about the distribution $Q^{\mathrm{do}(X^k := \bar{f}(\bar{\mathrm{PA}}^k, \bar{N}^k))}$. We obtain similar expressions when intervening on several variables at the same time.

Similar distributional shifts can be obtained, of course, if the factorization is non-causal (see also Section 3.1), so while our framework contains intervention distributions as a special case, it equally well applies to non-causal models.

[1] For notational convenience we sometimes refer to the parent sets by their indices, i.e., $\mathrm{PA}^j \subseteq \{1, \dots, d\} \setminus \{j\}$.

A.2. Efficient computation of $V(n,m)$ in Theorem 4

In this section, we show that for $n, m \in \mathbb{N}$ and $K \geq 1$

$$V(n,m) = \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{l} \binom{n-m}{m-\ell} (K^\ell - 1)$$

can be evaluated efficiently. If $mn/2$, such that for some ℓ one has $m-\ell \geq n-m$, we use the convention that if $a > b$ then $\binom{b}{a} = 0$. If one evaluated the term $\binom{n}{m}^{-1}$ separately, this could potentially cause numerical underflow, and similarly terms in the sum could get very large, such as the summand including $K^m - 1$.

Denote the summands by s_ℓ, that is

$$s_\ell = \frac{\binom{m}{l}\binom{n-m}{m-\ell}(K^\ell - 1)}{\binom{n}{m}}.$$

We can compute s_1 by:

$$\begin{aligned} s_1 &= \frac{\binom{m}{1}\binom{n-m}{m-1}}{\binom{n}{m}}(K-1) \\ &= m^2(K-1)\frac{(n-m)!(n-m)!}{n!(n-2m+1)!} \\ &= (K-1)\frac{m^2}{n-m+1}\prod_{j=0}^{m-2}\frac{n-m-j}{n-j}. \end{aligned}$$

This can be evaluated in $O(m)$ time. Further, if $s_\ell \neq 0$, the ratio of two consecutive summands is

$$\begin{aligned} \frac{s_{\ell+1}}{s_\ell} &= \frac{\binom{m}{\ell+1}\binom{n-m}{m-\ell-1}(K^{\ell+1}-1)}{\binom{m}{\ell}\binom{n-m}{m-\ell}(K^\ell - 1)} \\ &= \frac{(m-\ell)^2}{(\ell+1)(n-2m+\ell+1)}\frac{K^{\ell+1}-1}{K^\ell - 1}, \end{aligned}$$

which for a given ℓ, can be evaluated in $O(1)$ time. Hence, we can compute $\sum_{\ell=1}^m s_\ell$, by first computing s_1, and for each ℓ, compute $s_{\ell+1} = \frac{s_{\ell+1}}{s_\ell} s_\ell$, as long as $s_\ell \neq 0$ (after which the remaining terms are 0). The overall computational cost of computing $V(n,m) = \sum_{\ell=1}^m s_\ell$ is thus $O(m)$.

A.2.1. Plotting the level upper bound

To illustrate how the bound in Theorem 4 depends on $K = \mathbb{E}_Q[r(X)^2]$, α_φ, m and n, we plot the level bound $\inf_{\delta \in (0,1)} \left(\frac{\alpha_\varphi}{1-\delta} + \frac{V(n,m)}{V(n,m)+\delta^2} \right)$, for various values of α_φ, n, m and K in Fig. A.1. By choosing a very small resample size (such as $m = n^{0.2}$) and a conservative

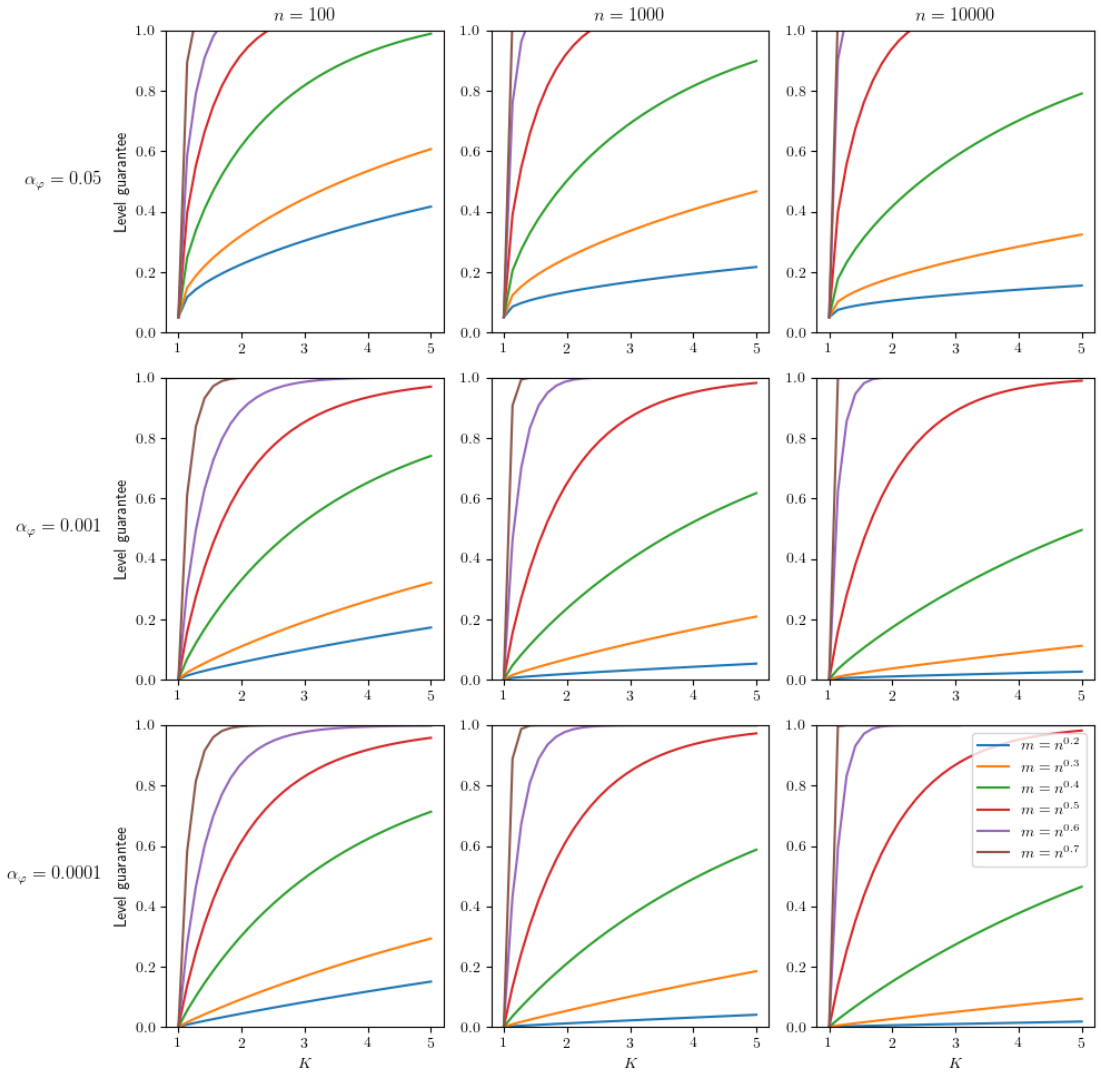


Figure A.1.: The finite level upper bound in Theorem 4 plotted for various values of m, n, α_φ and K. Here $K = \mathbb{E}_Q[r(X)^2]$ is a measure of the difference between the observed and the target domains and α_φ is the level of the target test applied.

target test (such as $\alpha_\varphi = 0.0001$), the level bound grows very slowly in K. For larger resample sizes m, the finite level is only guaranteed when K is close to 1, meaning that P^* and Q^* are similar. As we note in Section 4.4.1, the finite-level bound is in many cases weak, will lead to choices of m which reduces power, and we therefore in practice recommend to use the GOF-heuristic (Section 4.4.1) in combination with the procedure for combining multiple resamples (Section 4.4.2) instead.

A.3. Algorithm for hypothesis testing with unknown distributional shift

This section contains Algorithm A.1, which describes our method for testing under distributional shifts for the case where the shift factor r_q is unknown, but can be estimated by an estimator $\hat{r}$. Algorithm A.1 is similar to Algorithm 1 but one additionally splits the sample $\mathbf{X}_n$ into two disjoint samples $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ and uses $\mathbf{X}_{n_1}$ for estimating the weights $\hat{r}_{n_1}$, which are then, together with $\mathbf{X}_{n_2}$, used as an input to Algorithm 1.

We view the sample splitting as a theoretical device. In practice, we are using the full sample both for estimating the weights and for applying the test.

Algorithm A.1 Testing a target hypothesis with unknown distributional shift and resampling

Input: Data $\mathbf{X}_n$, target sample size m, hypothesis test φ_m, estimator $\hat{r}$ for r_q, and a.

1: Let n_1, n_2 be s.t. $n_1 + n_2 = n$ and $n_1^a = \sqrt{n_2}$
2: $\mathbf{X}_{n_1} \leftarrow X_1, \ldots, X_{n_1}$
3: $\mathbf{X}_{n_2} \leftarrow X_{n_1+1}, \ldots, X_{n_1+n_2}$
4: $\hat{r}_{n_1} \leftarrow$ estimate of r_q based on on $\mathbf{X}_{n_1}$
5: $\Psi_{\texttt{DRPL}}^{\hat{r}_{n_1},m}(\mathbf{X}_{n_2}, U) \leftarrow (X_{n_1+i_1}, \ldots, X_{n_1+i_m})$

return $\psi_n^{\hat{r}}(\mathbf{X}_n, U) \coloneqq \varphi_m(\Psi_{\texttt{DRPL}}^{\hat{r}_{n_1},m}(\mathbf{X}_{n_2}, U))$

A.4. Sampling from $\Psi_{\texttt{DRPL}}$

This section provides details on sampling from $\Psi_{\texttt{DRPL}}^{r,m}$, as defined by (10). We have defined $\Psi_{\texttt{REPL}}^{r,m}$ and $\Psi_{\texttt{NO-REPL}}^{r,m}$ as weighted resampling with and without replacement, respectively. $\Psi_{\texttt{NO-REPL}}^{r,m}$ can be implemented as a sequential procedure that first draws i_1 with weights $r(X_i)/\sum_{j=1}^n r(X_j)$, and then draws i_2 with weights $r(X_i)/\sum_{j=1,j\neq i_1}^n r(X_j)$, and so forth. Although both $\Psi_{\texttt{NO-REPL}}^{r,m}$ and $\Psi_{\texttt{DRPL}}^{r,m}$ sample distinct sequences $(i_1, \ldots, i_m)$, they are, in general, not equivalent, as can be seen from the form of the weights $w_{(i_1,\ldots,i_m)}^{\texttt{DRPL}}$ and $w_{(i_1,\ldots,i_m)}^{\texttt{NO-REPL}}$ below. When there is no ambiguity, we omit superscripts and write $\Psi_{\texttt{DRPL}}$, for example. We also interchangeably consider a sample from $\Psi_{\texttt{DRPL}}$ to be a sequence $(i_1, \ldots, i_m)$ and a subsample $(X_{i_1}, \ldots, X_{i_m})$ of $\mathbf{X}_n$.

The procedures $\Psi_{\texttt{DRPL}}$, $\Psi_{\texttt{REPL}}$ and $\Psi_{\texttt{NO-REPL}}$ sample a sequence $(i_1, \ldots, i_m)$ with weights

$w_{(i_1,\dots,i_m)}$ that are, respectively, given by:

$$w^{\texttt{DRPL}}_{(i_1,\dots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})} \quad \text{for distinct } (i_1,\dots,i_m)$$

$$w^{\texttt{REPL}}_{(i_1,\dots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{(j_1,\dots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} \quad \text{for all } (i_1,\dots,i_m)$$

$$w^{\texttt{NO-REPL}}_{(i_1,\dots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{j_1=1}^n r(X_{j_1}) \sum\limits_{\substack{j_2=1\\ j_2\neq i_1}}^n r(X_{j_2}) \cdots \sum\limits_{\substack{j_m=1\\ j_m\notin\{i_1,\dots,i_{m-1}\}}}^n r(X_{j_m})} \quad \text{for distinct } (i_1,\dots,i_m)$$

Here, the comment 'for distinct $(i_1,\dots,i_m)$' implies that the weights are zero otherwise. Most statistical software have standard implementations for sampling from $\Psi_{\texttt{REPL}}$ and $\Psi_{\texttt{NO-REPL}}$ (known simply as sampling with or without replacement). We now detail a number of ways to sample a sequence $(i_1,\dots,i_m)$ from $\Psi_{\texttt{DRPL}}$. The first two sampling methods are exact, the third sampling method is approximate.

A.4.1. Acceptance-rejection sampling with $\Psi_{\texttt{REPL}}$ as proposal

Given a sample $\mathbf{X}_n$, one can sample from $\Psi_{\texttt{DRPL}}$ by acceptance-rejection sampling from $\Psi_{\texttt{REPL}}$, by drawing sequences $(i_1,\dots,i_m)$ from $\Psi_{\texttt{REPL}}$ until one gets a draw that is distinct, which is then used as the draw from $\Psi_{\texttt{DRPL}}$. This is a valid sampling method for $\Psi_{\texttt{DRPL}}$, because for any distinct sequence $(i_1,\dots,i_m)$ we have

$$\begin{aligned}
\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\dots,i_m) \mid \Psi_{\texttt{REPL}} \text{ distinct}, \mathbf{X}_n) &= \frac{\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\dots,i_m), \Psi_{\texttt{REPL}} \text{ distinct} \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{REPL}} \text{ distinct} \mid \mathbf{X}_n)} \\
&= \frac{\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\dots,i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{REPL}} \text{ distinct} \mid \mathbf{X}_n)} \\
&= \frac{w^{\texttt{REPL}}_{(i_1,\dots,i_m)}}{\sum\limits_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} w^{\texttt{REPL}}_{(j_1,\dots,j_m)}} \\
&= \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{(j_1,\dots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} \frac{\sum\limits_{(j_1,\dots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})}{\sum\limits_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})} \\
&= w^{\texttt{DRPL}}_{(i_1,\dots,i_m)} \\
&= \mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1,\dots,i_m) \mid \mathbf{X}_n).
\end{aligned}$$

By integrating over $\mathbf{X}_n$, this implies that $\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\dots,i_m) \mid \Psi_{\texttt{REPL}} \text{ distinct}) = \mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1,\dots,i_m))$. Proposition A.1 shows that under certain assumptions, the

probability of sampling a distinct sample from $\Psi_{\texttt{REPL}}$ converges to 1.

A.4.2. Acceptance-rejection sampling with $\Psi_{\texttt{NO-REPL}}$ as proposal

If m is large compared to n, it may be that most of the samples drawn from $\Psi_{\texttt{REPL}}$ are not distinct, and so the acceptance rejection scheme in Appendix A.4.1 may take too many attempts to produce a distinct sample. As an alternative, one can use $\Psi_{\texttt{NO-REPL}}$ as a proposal distribution for an acceptance-rejection sampler, which is typically faster, since $\Psi_{\texttt{NO-REPL}}$ has the same support as $\Psi_{\texttt{DRPL}}$. Given a sample $\mathbf{X}_n$, we thus need to identify an M such that

$$\forall \text{ distinct } (i_1, \ldots, i_m) : \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)} \leq M.$$

We have

$$\frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)} = \frac{w^{\texttt{DRPL}}_{(i_1,\ldots,i_m)}}{w^{\texttt{NO-REPL}}_{(i_1,\ldots,i_m)}} = \frac{\sum_{j_1} r(X_{j_1}) \sum_{j_2 \neq i_1} r(X_{j_2}) \cdots \sum_{j_m \neq i_1,\ldots,i_{m-1}} r(X_{j_m})}{\sum_{\substack{(j_1,\ldots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}.$$

The denominator does not depend on $i_1, \ldots, i_m$, and the numerator can be upper bounded by

$$(1-0)(1-p_{(1)})(1-p_{(1)}-p_{(2)}) \cdots (1-p_{(1)}-\ldots-p_{(m-1)}),$$

where $p_{(1)} = \min\{r(X_1), \ldots, r(X_n)\}$ is the smallest of the weights, $p_{(2)}$ is the second smallest, etc. Thus, we can choose

$$M := \frac{(1-0)(1-p_{(1)})(1-p_{(1)}-p_{(2)}) \cdots (1-p_{(1)}-\ldots-p_{(m-1)})}{\sum_{\substack{(j_1,\ldots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}.$$

We now proceed with an ordinary acceptance-rejection sampling scheme: We sample a (distinct) sequence $(i_1, \ldots, i_m)$ from $\Psi_{\texttt{NO-REPL}}$ and an independent, uniform variable V on the interval $(0, 1)$. We accept $(i_1, \ldots, i_m)$ if

$$\begin{aligned} V &\leq \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)}{M \cdot \mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)} \\ &= \frac{(1-0)(1-p_{i_1})(1-p_{i_1}-p_{i_2}) \cdots (1-p_{i_1}-\ldots-p_{i_{m-1}})}{(1-0)(1-p_{(1)})(1-p_{(1)}-p_{(2)}) \cdots (1-p_{(1)}-\ldots-p_{(m-1)})}. \end{aligned}$$

Here, we have used that the denominator of M cancels with the normalization constant of $\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)$. If the sample is not accepted, we draw another sample from $\Psi_{\texttt{NO-REPL}}$ until one sample is accepted.

A.4.3. Approximate Gibbs sampling starting from $\Psi_{\texttt{NO-REPL}}$

There are cases, where the sampling schemes presented in Appendices A.4.1 and A.4.2 do not yield an accepted sample in a reasonable amount of time (this is typically due to m being too large compared to n, see Assumption (A1)). In such cases, one can get an approximate sample of $\Psi_{\texttt{DRPL}}$ by sampling $\Psi_{\texttt{NO-REPL}}$ and shifting it towards $\Psi_{\texttt{DRPL}}$ using a Gibbs sampler [Geman and Geman, 1984].

Let therefore $(i_1, \dots, i_m)$ be an initial (distinct) sample from $\Psi_{\texttt{NO-REPL}}$, and define $i_{-\ell}$ to be the sequence without the ℓ'th entry. The Gibbs sampler sequentially samples i_ℓ from the conditional distribution $j \mid i_{-\ell}$ in $\Psi_{\texttt{DRPL}}$. To compute this conditional probability let $\Psi^\ell_{\texttt{DRPL}}$ be the ℓ'th index of a sample. Then

$$\begin{aligned}
&\mathbb{P}_Q(\Psi^\ell_{\texttt{DRPL}} = j | \Psi^{-\ell}_{\texttt{DRPL}} = i_{-\ell}) \\
&\qquad = \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \dots, j, \dots, i_m))}{\mathbb{P}_Q(\Psi^{-\ell}_{\texttt{DRPL}} = i_{-\ell})} \\
&\qquad = \frac{r(X_{i_1}) \cdots r(X_j) \cdots r(X_{i_m})}{\sum_{v \notin i_{-\ell}} r(X_{i_1}) \cdots r(X_v) \cdots r(X_{i_m})} = \frac{r(X_j)}{\sum_{v \notin i_{-\ell}} r(X_v)},
\end{aligned}$$

i.e., the conditional distribution of one index i_ℓ given $i_{-\ell}$ is just a weighted draw among $\{i_1, \dots, i_m\} \backslash i_{-\ell}$. This is simple to sample from and the Gibbs sampler now iterates through the indices $(i_1, \dots, i_m)$, at each iteration replacing the index i_ℓ by a sample from the conditional given $i_{-\ell}$. Iterating this a large number of times produces an approximate sample from $\Psi_{\texttt{DRPL}}$.

A.5. Sampling with replacement

Instead of the sampling scheme $\Psi_{\texttt{DRPL}}$ presented above, we can also use weighted sampling with replacement, which we denote $\Psi_{\texttt{REPL}}$ (see Appendix A.4 for details). Sampling from $\Psi_{\texttt{REPL}}$ is simpler than from $\Psi_{\texttt{DRPL}}$, and while sampling from $\Psi_{\texttt{REPL}}$ is in some cases disadvantageous for testing (e.g., if we test whether the target distribution has a point mass), if the test is not prone to duplicate data points, testing based on $\Psi_{\texttt{REPL}}$ may be advantageous over $\Psi_{\texttt{DRPL}}$ (further examination is needed to clarify this relationship). When sampling without weights, Bickel et al. [2012] present regularity conditions on the test statistic that guarantee consistency even with $m = o(n)$.

Here we show that under additional assumptions, the probability of a non-distinct sample from $\Psi_{\texttt{REPL}}$ converges to 0. Consider the following strengthening of Assumption (A2).

(A3) There exists $L \in \mathbb{R}$ such that for all $v \geq 1$, $\mathbb{E}_Q[r(X_i)^{v+1}] \leq L^v$.

This is for instance trivially satisfied if $r(X_i)$ is Q-a.s. bounded by a constant L. The following proposition shows that under Assumptions (A1) and (A3) the probability of drawing a distinct sample from $\Psi_{\texttt{REPL}}$ converges to 1.

Proposition A.1 (Asymptotic equivalence of REPL and DREPL for bounded weights)**.** *Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r : \mathcal{X} \to [0, \infty)$ exists, satisfying $\tau(q)(x) \propto r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let $m = m(n)$ be a resampling size and let Ψ_{REPL} be the weighted resampling with replacement defined in Section 4.1. Then, if m and Q satisfy Assumptions (A1) and (A3), it holds that*

$$\lim_{n\to\infty} \mathbb{P}_Q(\Psi^{r,m}_{REPL}(\mathbf{X}_n, U) \text{ distinct}) = 1.$$

As a corollary to Theorem 1, we also have pointwise asymptotic level of a test when $\Psi_{\texttt{REPL}}$ is used instead of $\Psi_{\texttt{DRPL}}$.

Corollary A.1 (Pointwise asymptotics – REPL)**.** *Assume the same setup and assumptions as in Theorem 1 and additionally assume Assumption (A3). Let Ψ_{REPL} be the weighted resampling with replacement defined in Section 4.1 and let ψ^r_n be the REPL-based resampling test defined by $\psi^r_n(\mathbf{X}_n, U) := \varphi_m(\Psi^{r,m}_{REPL}(\mathbf{X}_n, U))$. Then, it holds that*

$$\limsup_{n\to\infty} \mathbb{P}_Q(\psi^r_n(\mathbf{X}_n, U) = 1) = \alpha_\varphi.$$

The same statement holds when replacing both lim sup *'s (including the one in α_φ) with* lim inf *'s.*

A.6. Algorithm for the GOF-heuristic for choosing m

The GOF-heuristic is a data driven procedure to choose m in finite sample settings, see Section 4.4.1. It is summarized in Algorithm A.2.

Algorithm A.2 GOF-heuristic: Choosing m by testing resampling validity

Input: Data $\mathbf{X}_n$, shift factor $r(x^A)$, threshold α_c, initial target size m_0, increment size Δ, repetitions K, conditional goodness-of-fit test κ.

$\texttt{qt} \leftarrow \alpha_c$-quantile of $\texttt{mean}(U_1, \ldots, U_K)$, where $U_i \sim Unif(0, 1)$
$m \leftarrow m_0$
$\texttt{m\_valid} \leftarrow \texttt{true}$
while $\texttt{m\_valid}$ **do**
  for $k = 1, \ldots K$ **do**
    $\texttt{res}_k \leftarrow \kappa(\Psi^{r,m}(\mathbf{X}_n))$
  if $\texttt{mean}(\texttt{res}_1, \ldots, \texttt{res}_K) > \texttt{qt}$ **then**
    $m \leftarrow m + \Delta$
  else
    $\texttt{m\_valid} \leftarrow \texttt{false}$
    $m \leftarrow m - \Delta$
return m

A.7. Additional experiments

A.7.1. Assumption Assumption (A1) when sample size increases

As indicated by the results in Section 5.1, Assumption (A1) may sometimes be too strict of an assumption, in the sense that level can be attained also when Assumption (A1) is violated. We explored this for a continuous data example in Section 5.1, but here, we investigate the effect of violating Assumption (A1) as the sample size n increases for a binary setting.

We simulate data $\mathbf{X}_n$ from a binary distribution $\mathbb{P}_Q(X = 0) = 0.9$, $\mathbb{P}_Q(X = 1) = 0.1$, and consider the target distribution P where the probabilities are flipped , i.e. $\mathbb{P}_P(X = 0) = 0.1$, $\mathbb{P}_P(X = 1) = 0.9$. We consider the hypothesis $H_0 : \mathbb{P}_P(X = 1) > 0.8$, which is clearly satisfied for P, but may not be detected by the resample if $m = n^a$ for $a > 0.5$ chosen too large. We repeat the experiment 200 times and compute the resulting rejection rates for various sample sizes and rates $m = n^a$. The results are shown in Fig. A.2. Under Assumption (A1), Theorem 1 guarantees asymptotic level. Indeed, at all sample sizes n, the test has the correct level, when Assumption (A1) is satisfied (that is, when $a < 0.5$). Yet, though this is not guaranteed by theory, we have indications of asymptotic level also for $a > 0.5$.

A.7.2. Model selection under covariate shift

In this section, we apply our testing method to the problem of model selection under covariate shift as discussed in Section 3.6. We generate a data set $D := \{(X_i, Y_i)\}_{i=1}^n$ of size $n = 3'000$. Each (X_i, Y_i) is drawn i.i.d. according to the following data generating process:

$$(X^1, X^2) \sim \texttt{GaussianMixture}(\begin{bmatrix}3 & 3\end{bmatrix}, \begin{bmatrix}-3 & -3\end{bmatrix}) \qquad Y := \begin{cases}\sin(X_2 + \varepsilon_Y), & \text{if } X_1 \geq 0\\ \sin(3 + X_1 + \varepsilon_Y), & \text{if } X_1 < 0,\end{cases}$$

where $\texttt{GaussianMixture}(\begin{bmatrix}3 & 3\end{bmatrix}, \begin{bmatrix}-3 & -3\end{bmatrix})$ is an even mixture (i.e., $p = 0.5$) of two 2-dimensional Gaussian distributions with means $\mu_1 = (3, 3)^\top$, $\mu_2 = (-3, -3)^\top$ and unit covariance matrix, and ε_Y is a standard Gaussian $\mathcal{N}(0, 1)$-variable. Fig. A.3 illustrates a sample of size $1'000$ from this data generating process.

We randomly split the data D into a training set D_{train} of size $2'000$ and a test set D_{test} of size $1'000$. Using D_{train}, we train two candidate classifiers, namely logistic regression (`LR`) and random forest (`RF`) to predict Y from X. Both models are trained using the `Scikit-Learn` Python package [Pedregosa et al., 2011] with default parameters. We consider the area under the curve (AUC) as the scoring function, where we denote the AUC scores for the models `LR` and `RF` by AUC(`LR`) and AUC(`RF`), respectively. Then, we apply our resampling approach on D_{test} to test whether `LR` outperforms `RF` when the distribution of (X^1, X^2) is changed to a single 2-dimensional Gaussian distribution with mean $\mu = (3, 3)^\top$ and unit covariance matrix. In this experiment, we choose the resampling size m by the GOF-heuristic (see Algorithm A.2) and assume that the shift factor $r(x) := p^*(x)/q^*(x)$ is known, where $q^*(x)$ is a pdf of the mixture

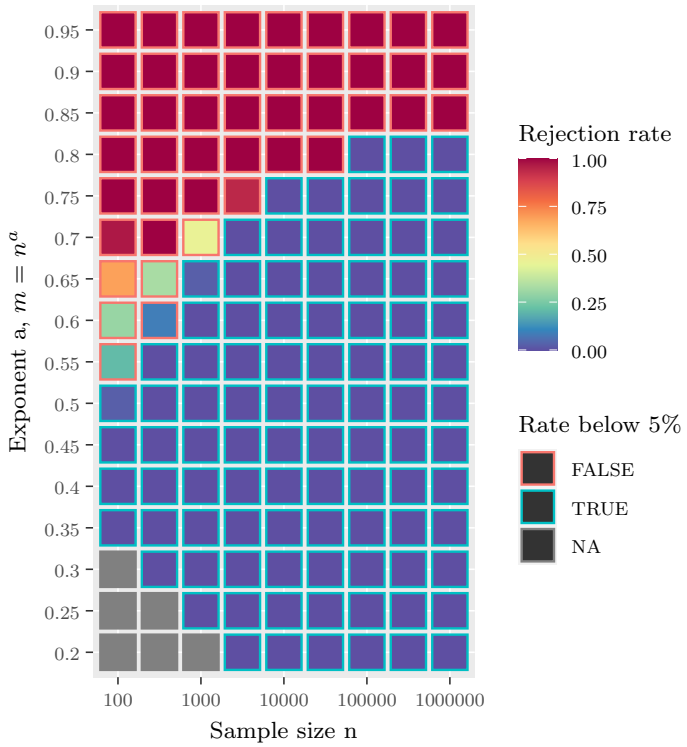


Figure A.2.: Rejection rates when $m = n^a$ points are resampled for various exponents a and sample sizes n. Each tile represents a combination of sample size n and exponent a, and the color indicates the rate at which the hypothesis $\mathbb{P}_P(X = 1) > 0.8$ is rejected in the resampled data, when resampling $m = n^a$ points; ideally this should be low (blue) since we consider a target distribution where $\mathbb{P}_P(X = 1) = 0.9$. Even though this is not provided by our theoretical results, it seems that level is possible even for rates larger than 0.5.

of Gaussian $\texttt{GaussianMixture}(\begin{bmatrix}3 & 3\end{bmatrix}, \begin{bmatrix}-3 & -3\end{bmatrix})$ and $p^*(x)$ is a pdf of the Gaussian $\mathcal{N}((3,3)^\top, \boldsymbol{I}_2)$. We employ DeLong's test [DeLong et al., 1988] to test the hypothesis $\text{AUC}(\texttt{LR}) \leq \text{AUC}(\texttt{RF})$ using the R-package `pROC` [Robin et al., 2011].

We repeat the experiment 500 times and report in Table A.1 how many times our resampling test (resampling) rejects and returns that $\text{AUC}(\texttt{LR}) > \text{AUC}(\texttt{RF})$. As a benchmark, we also report the corresponding rejection rate when we perform the test directly on a sample from the target distribution (oracle) in which the covariate distribution is changed to $p^*(x)$. We also report the rejection rate when we perform the test on the observed test set directly D_{test} (observed), without resampling it first.

As shown in Table A.1, under the observed distribution we have $\text{AUC}(\texttt{LR}) \leq \text{AUC}(\texttt{RF})$ (the rejection rate is 0 in the observed sample). However, under the target distribution $\text{AUC}(\texttt{LR})$ is higher than $\text{AUC}(\texttt{RF})$ (the rejection rate is 1 in the target sample). Our resampling approach yields high power against the alternative hypothesis, even without

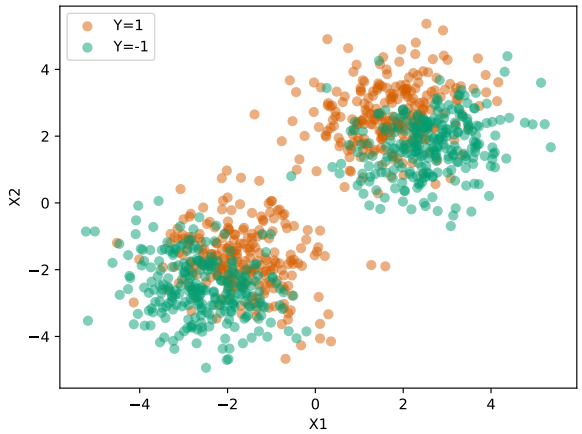


Figure A.3.: Sample from the data generating process for the experiment described in Appendix A.7.2.

resampling	oracle	observed
0.874 ± 0.029	1.0	0.0

Table A.1.: Rejection rates of the test with alternative AUC(`LR`) > AUC(`RF`) over 500 repetitions. The resampling test shows high power to detect the alternative.

having access to the oracle information but using resampling instead.

A.7.3. Conditional independence testing in the bikeshare dataset

Following Berrett et al. [2020], we consider the capital bikeshare dataset [Fanaee-T and Gama, 2014]. Berrett et al. [2020] test conditional independence of the duration X from three different outcomes Y_j, $j = 1, 2, 3$ where Y_1 is a binary 'User type' variable indicating whether the user is a member, Y_2 is the date and Y_3 is the day of the week, conditioning on a variable Z which encodes mean and variance for the particular ride and at that time of the day. More specifically, they use their proposed permutation test (CPT) and the randomization test (CRT) by Candès et al. [2018]. Here, we apply our method for testing conditional independence from Section 3.1 to the same data, using the GOF-heuristic (Section 4.4.1) to chose m. Since both our and their method rely on knowing the conditional $q(x|z)$, we can use their estimated conditionals $\hat{q}(x|z)$ such that differences in outcome is not because of differences in estimation.

We display the resulting p-values for the hypothesis of conditional independence in Table A.2. At a 5% significance level, our method rejects the same hypotheses as they do, finding that the duration is not conditionally independent of the user type, but is so of both the date and the day of the week.

Variable	CPT	CRT	Ours (combination test)	Our (single test)
User type	0.0010	0.0010	$1.447 \cdot 10^{-16}$	$4.595 \cdot 10^{-5}$
Date	0.1146	0.1293	0.1320	0.0586
Day of week	0.1980	0.2063	0.4395	0.1246

Table A.2.: Comparison of p-values for the conditional independence hypothesis $X \perp\!\!\!\perp Y_j | Z$ for three different variables Y_j. We compare to CPT [Berrett et al., 2020] and CRT [Candès et al., 2018] to our method using a single test or the combination test of Hartung [1999] as discussed in Section 4.4.2.

A.7.4. Additional simulation details

In this section, we state the models used for generating the data in Sections 5.2, 5.6, 5.9 and 5.10.

Section 5.2

The linear data in Section 5.2 was generated by the equations

$$Z := \varepsilon_Z \qquad X := Z + 2\varepsilon_X \qquad Y = Z + \theta X + 2\varepsilon_Y$$

and the nonlinear data was generated by the equations

$$X := \varepsilon_X \qquad Z := |X - 1| + 4\varepsilon_Z \qquad Y = Z + \theta X + 2\varepsilon_Y$$

where $\varepsilon_Z, \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$, and either $\theta = 0$ (no effect present) or $\theta = 0.5$ (effect present).

Section 5.6

The binary data in Section 5.6 was generated by first sampling hyper-parameters:

$$p_H \sim \text{Dirichlet}^{1\times 4}(3,3,3,3) \qquad p_1 \sim B^{1\times 1}(1,1) \qquad p_2 \sim B^{2\times 4}(1,1) p_3 \sim B^{1\times 2}(1,1)$$
$$\mathcal{G} : p_4 \sim B^{2,1}(1,1) \qquad \mathcal{H} : p_4 \sim B^{2,2}(1,1),$$

where $B(a, b)$ is a Beta distribution and the superscript indicates the dimension of the sampled parameter matrix. The matrices $p_1, \ldots, p_4$ correspond to conditional probability tables given parent variables in the graphs $\mathcal{G}$ or $\mathcal{H}$. The distributions are the same when sampling from $\mathcal{G}$ and $\mathcal{H}$, except for X^4, which has an additional parent in $\mathcal{H}$. In each repetition, given hyper-parameters, a data set is sampled from the structural equation

model

$$H \sim \text{choice}(\{1,\dots,4\}, \text{weights} = p_H) \qquad X^1 \sim \text{Bernoulli}(p_1)$$
$$X^2 \sim \text{Bernoulli}(p_{2_{X^1,H}}) \qquad X^3 \sim \text{Bernoulli}(p_{3_{X^2}})$$
$$\mathcal{G}: X^4 \sim \text{Bernoulli}(p_{4_{X^3}}) \qquad \mathcal{H}: X^4 \sim \text{Bernoulli}(p_{4_{X^3,X^1}})$$

where the subscript $p_{2_{X^1,U}}$ indicate that for an outcome of (X^1, U), the Bernoulli distribution uses the probability in the corresponding entry of p_2 (and similar for p_3 and p_4).

The Gaussian data in Section 5.6 was generated by the structural equation model

$$H \coloneqq \varepsilon_H \qquad X^1 \coloneqq \varepsilon_{X^1} \qquad X^2 \coloneqq X^1 + H + \varepsilon_{X^2}$$
$$X^3 \coloneqq X^2 + 2\varepsilon_{X^3} \qquad X^4 \coloneqq \theta \cdot X^1 + X^3 + H + \varepsilon_{X^4}$$

where $\varepsilon_H, \varepsilon_{X^j} \sim \mathcal{N}(0,1)$, and $\theta \in \{0, 0.3\}$ indicates the absence or presence of the edge $X^1 \to X^4$.

The non-Gaussian data in Section 5.6 was generated by the structural equation model

$$H \coloneqq \tfrac{1}{2} \cdot \varepsilon_H \cdot \varepsilon_H \qquad X^1 \coloneqq \gamma_{X^1} \qquad X^2 \coloneqq X^1 \cdot H + \varepsilon_{X^2}$$
$$X^3 \coloneqq X^2 \cdot X^2 + \tfrac{3}{2}\varepsilon_{X^3} \qquad X^4 \coloneqq \theta \cdot X^1 + X^3 + H + \varepsilon_{X^4},$$

where $\varepsilon_H, \varepsilon_{X^j} \sim \mathcal{N}(0,1)$ and γ_{X^1} follows a $\Gamma(2)$-distribution, and $\theta \in \{0, 0.3\}$ indicates the absence or presence of the edge $X^1 \to X^4$.

Section 5.9

The data in Section 5.9 was generated by the structural equation model

$$X^1 \coloneqq \varepsilon_{X^1} \qquad X^3 \coloneqq \varepsilon_{X^3} \qquad X^2 \coloneqq X^3 + X^1 + 2\varepsilon_{X^2}$$
$$Y \coloneqq X^2 + X^3 + 0.3\varepsilon_Y \qquad X^4 \coloneqq -Y + X^2 + X^3 + 0.7\varepsilon_{X^4}$$

Section 5.10

The linear data in Section 5.10 was generated by the equations

$$X \coloneqq \varepsilon_X \qquad \mathbb{P}(D=1) = \sigma\left(\frac{2}{\sqrt{20}}\sum_{i=1}^{20} X_i\right) \qquad Y \coloneqq \theta D + (1+\theta D)\left(\frac{1}{5}\sum_{i=1}^{20} X_i\right) + 0.3\varepsilon_Y$$

and the non-linear data was generated by the equations

$$X \coloneqq \varepsilon_X \qquad \mathbb{P}(D=1) = \sigma\left(\frac{2}{\sqrt{20}}\sum_{i=1}^{20} X_i\right)$$

$$Y \coloneqq \theta D + (1+\theta D)\left(0.2\sum_{i=1}^{20} X_i + \exp\left(-0.2\left(\sum_{i=1}^{20} X_i\right)^2/2\right) * \sin(0.2\sum_{i=1}^{20} X_i)\right) + 0.3\varepsilon_Y$$

where $\varepsilon_X \sim \mathcal{N}(0, \mathrm{Id}_{20})$ and $\varepsilon_Y \sim \mathcal{N}(0,1)$, and either $\theta = 0$ (no effect present) or $\theta = 0.3$ (effect present).

A.8. Proofs

A.8.1. Proof of Theorem 1

Proof of Theorem 1. We show the statement only for lim sup, the corresponding statement for lim inf follows by replacing lim sup with lim inf everywhere.

Let p and q denote the respective densities of P and Q with respect to the dominating measure μ. By assumption $p = \tau(q)$, so $p(x) \propto r(x)q(x)$. Let $\bar{r}$ be the normalized version of r satisfying $p(x) = \bar{r}(x)q(x)$. Recall that we call a sequence $(i_1, \ldots, i_m)$ distinct if for all $\ell \neq \ell'$ we have $i_\ell \neq i_{\ell'}$. The resampling scheme $\Psi_{\texttt{DRPL}}$, defined by (10), samples from the space of distinct sequences $(i_1, \ldots, i_m)$, where every sequence has probability $w_{(i_1,\ldots,i_m)} \propto \prod_{\ell=1}^m r(X_{i_\ell})$. The normalization constant here is the sum over the weights in the entire space of distinct sequences, that is,

$$w_{(i_1,\ldots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})} = \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})}.$$

Thus, taking an expectation involving $\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$, amounts to evaluating φ_m in all distinct sequences $X_{i_1}, \ldots, X_{i_m}$ and weighting with the probabilities $w_{(i_1,\ldots,i_m)}$.

$$\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1) = \mathbb{E}_Q\left[\frac{\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}}}{\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}}\prod_{\ell=1}^m \bar{r}(X_{j_\ell})}\right], \tag{A.1}$$

where we divide by the number of distinct sequences $\frac{n!}{(n-m)!}$ in both numerator and denominator.

Let $c(n,m)$ and $d(n,m)$ be the numerator and denominator terms of (A.1), i.e.,

$$c(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \left(\prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) \right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}},$$

$$d(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{j_\ell}).$$

We want to show that $\limsup_{n\to\infty} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] = \alpha_\varphi$. To see this, define for all $\delta > 0$ the set $A_\delta := \{|d(n,m) - 1| \le \delta\}$. It holds for all $\delta \in (0,1)$ that

$$\begin{aligned}
\mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] &= \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\le \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\le \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\right] + \mathbb{P}_Q\left(A_\delta^c\right) \\
&= \frac{1}{1-\delta}\mathbb{P}_P(\varphi_m(X_1,\dots,X_m)=1) + \mathbb{P}_Q\left(A_\delta^c\right),
\end{aligned}$$

where we used that $\frac{c(n,m)}{d(n,m)} \le 1$ and Lemma A.1 (a). Further combining Chebyshev's inequality with Lemma A.1 (b) and (d), it follows that

$$\lim_{n\to\infty} \mathbb{P}_Q\left(A_\delta^c\right) \le \lim_{n\to\infty} \frac{\mathbb{E}_Q[(d(n,m)-1)^2]}{\delta^2} = \lim_{n\to\infty} \frac{\mathbb{V}_Q(d(n,m))}{\delta^2} = 0.$$

Hence, using that $\limsup_{k\to\infty} \mathbb{P}_P(\varphi_k(X_1,\dots,X_k)=1) = \alpha_\varphi$ we have shown for all $\delta \in (0,1)$ that

$$\limsup_{n\to\infty} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] \le \frac{1}{1-\delta}\alpha_\varphi. \tag{A.2}$$

Similarly, we also get for all $\delta \in (0,1)$ the following lower bound

$$\begin{aligned}
\mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] &= \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\ge \mathbb{E}_Q\left[\frac{c(n,m)}{1+\delta}\mathbb{1}_{A_\delta}\right] \\
&\ge \frac{1}{1+\delta}\mathbb{E}_Q\left[c(n,m)\right],
\end{aligned}$$

where in the last inequality we used that $c(n,m) \ge 0$. Again using Lemma A.1 (a) and

that $\limsup_{k\to\infty} \mathbb{P}_P(\varphi_k(X_1,\dots,X_k)=1) = \alpha_\varphi$, we get that for all $\delta \in (0,1)$ that

$$\limsup_{n\to\infty} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] \geq \frac{1}{1+\delta}\alpha_\varphi. \tag{A.3}$$

using that $\delta \in (0,1)$ is arbitrary, this proves that

$$\limsup_{n\to\infty} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] = \alpha_\varphi,$$

which completes the proof of Theorem 1.

□

Remark A.1. After Theorem 2, we discussed the conjecture that for any $\ell \geq 2$ we can relax Assumption (A1) to $m = o(n^{1-1/\ell})$ if one changes Assumption (A2) to $\mathbb{E}_Q[r(X_i)^\ell] < \infty$. We now argue why the current structure of the proof of Theorem 1 does not allow for proving this statement. In the proof, we use that $\mathbb{V}_Q[d(n,m)]$ converges to 0, and the expression for the variance for U-statistics that we use in (A.7) in Lemma A.1 only depends on the second moment of the weights. In particular, the vanishing of the variances only depends on the relation between m and n, and not higher order moments of $\mathbb{E}[r(X)^\ell]$, and even if those were finite, the counterexample in Theorem 2 shows that the variance of $c(n,m)$ and $d(n,m)$ would still approach infinity if $m \neq o(\sqrt{n})$ i.e., m grows at least as fast as $\sqrt{n}$.

Thus, if one were to prove a result for the case $\mathbb{E}_Q[r(X_i)^\ell] < \infty$, it appears to us that one needs to change the proof of Theorem 1 to treat $\mathbb{E}_Q[c(n,m)/d(n,m)]$ jointly rather than using Lemma A.1 to treat each of $c(n,m)$ and $d(n,m)$ separately.

Lemma A.1 (Distinct draws)**.** *Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ have densities p and q with respect to a dominating measure μ. Let $\bar{r} : \mathcal{X} \to [0,\infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) = \bar{r}(x)q(x)$. Let $c(n,m)$ and $d(n,m)$ be defined by*

$$c(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\dots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}, \tag{A.4}$$

$$d(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(j_1,\dots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m \bar{r}(X_{j_\ell}). \tag{A.5}$$

Then, if m and Q satisfy Assumption (A1) and Assumption (A2) it holds that

(a) $\mathbb{E}_Q[c(n,m)] = \mathbb{P}_P(\varphi_m(X_1,\dots,X_m)=1)$,

(b) $\mathbb{E}_Q[d(n,m)] = 1$,

(c) $\lim_{n\to\infty} \mathbb{V}_Q[c(n,m)] = 0$,

(d) $\lim_{n\to\infty} \mathbb{V}_Q[d(n,m)] = 0$.

Proof. (A) we first prove the statements for the means, i.e., (a) and (b), and (B) we then prove the statements for the variances, i.e., (c) and (d).

Part A (means): Define $\delta_m := \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}$ (for the case (A.4)) or $\delta_m := 1$ (for the case (A.5)). Then, in both cases it holds that

$$\begin{aligned}
&\mathbb{E}_Q\left[\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\delta_m\right] \\
&= \frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\mathbb{E}_Q\left[\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\delta_m\right] \\
&= \frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\int\left(\prod_{\ell=1}^m \bar{r}(x_{i_\ell})q(x_{i_\ell})\right)\delta_m \mathrm{d}\mu^m(x_{i_1},\dots,x_{i_m}) \\
&= \frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\int\left(\prod_{\ell=1}^m p(x_{i_\ell})\right)\delta_m \mathrm{d}\mu^m(x_{i_1},\dots,x_{i_m}) \\
&= \frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\mathbb{E}_P[\delta_m] \\
&= \mathbb{E}_P[\delta_m]
\end{aligned}$$

In the second and fourth equality, we use that $i_1,\dots,i_m$ are all distinct, and in the last equality, we use that the number of distinct sequences $(i_1,\dots,i_m)$ is $\frac{n!}{(n-m)!}$. Consequently the term in (A.4) has mean $\mathbb{E}_P[\mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}] = \mathbb{P}_P(\varphi_m(X_1,\dots,X_m)=1)$ and the term in (A.5) has mean 1.

Part B (variances): We begin by expressing $\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\delta_m$ as a U-statistic [Serfling, 1980]. A U-statistic has the form

$$\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} h_m(Z_{i_1},\dots,Z_{i_m}) \tag{A.6}$$

for some symmetric function $h_m(z_1,\dots,z_m)$ (called a kernel function). In our case, the kernel function is $h_m(X_{i_1},\dots,X_{i_m}) := \prod_{\ell=1}^m \bar{r}(X_{i_\ell})\delta_m$. The variance of the correspond-

ing U-statistic [see Serfling, 1980, Section 5.2] is given by

$$\mathbb{V}_Q\left(\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell})\right)\delta_m\right)=\binom{n}{m}^{-1}\sum_{v=1}^{m}\binom{m}{v}\binom{n-m}{m-v}\zeta_v \tag{A.7}$$

where for all $v \in \{1,\ldots,m\}$

$$\zeta_v := \mathbb{V}_Q\left(\mathbb{E}_Q[h_m(X_{i_1},\ldots,X_{i_m}) \mid X_{i_1},\ldots,X_{i_v}]\right).$$

We now bound ζ_v from above by the second moment as follows

$$\zeta_v \leq \mathbb{E}_Q\left[\mathbb{E}_Q[h_m(X_{i_1},\ldots,X_{i_m}) \mid X_{i_1},\ldots,X_{i_v}]^2\right].$$

Moreover, using that δ_m is upper bounded by 1, we get for both cases (A.4) and (A.5) that

$$\begin{aligned}\zeta_v &\leq \mathbb{E}_Q\left[\mathbb{E}_Q[h_m(X_{i_1},\ldots,X_{i_m}) \mid X_{i_1},\ldots,X_{i_v}]^2\right]\\ &\leq \mathbb{E}_Q\left[\mathbb{E}_Q\left[\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell}) \mid X_{i_1},\ldots,X_{i_v}\right]^2\right].\end{aligned} \tag{A.8}$$

Next, since $(i_1,\ldots,i_m)$ are distinct, the variables $X_{i_1},\ldots,X_{i_m}$ are independent. Hence we have that

$$\begin{aligned}&\mathbb{E}_Q\left[\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell}) \mid X_{i_1},\ldots,X_{i_v}\right]\\ &=\left(\prod_{\ell=1}^{v}\bar{r}(X_{i_\ell})\right)\prod_{\ell=v+1}^{m}\mathbb{E}_Q\left[\bar{r}(X_{i_\ell})\right]\\ &=\left(\prod_{\ell=1}^{v}\bar{r}(X_{i_\ell})\right),\end{aligned} \tag{A.9}$$

where the last equality follows because

$$\mathbb{E}_Q[\bar{r}(X_{i_\ell})] = \int \bar{r}(x)q(x)\mathrm{d}\mu(x) = \int p(x)\mathrm{d}\mu(x) = 1.$$

Next, combining (A.8) and (A.9) we get that

$$\begin{aligned}\zeta_v &\leq \mathbb{E}_Q\left[\left(\prod_{\ell=1}^{v} \bar{r}(X_{i_\ell})\right)^2\right] \\ &= \prod_{\ell=1}^{v} \mathbb{E}_Q\left[\bar{r}(X_{i_\ell})^2\right] \\ &= \mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right]^v.\end{aligned}$$

Here, we again use the independence of the distinct terms. Plugging this into (A.7), we get

$$\begin{aligned}&\mathbb{V}_Q\left(\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \left(\prod_{\ell=1}^{m} \bar{r}(X_{i_\ell})\right) \delta_m\right) \\ &\leq \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} \mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right]^\ell.\end{aligned}$$

By Assumption (A2), $\mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right] < \infty$, so Lemma A.2 implies that this converges to 0 for $n \to \infty$. This shows that the variance converges to zero in both cases (A.4) and (A.5), which completes the proof of Lemma A.1. □

Lemma A.2. *Let $m = o(\sqrt{n})$ as n goes to infinity. Then for any $K \geq 0$, it holds that*

$$\lim_{n\to\infty} \frac{1}{\binom{n}{m}} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} K^\ell = 0. \tag{A.10}$$

Remark A.2. The Chu-Vandermonde identity states that $\frac{1}{\binom{n}{m}} \sum_{\ell=0}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} = 1$. In light of this identity, one may be surprised that when including the exponentially growing term, K^ℓ, the sum vanishes. The reason is that the summation in (A.10) starts at $\ell = 1$, not $\ell = 0$, and since n grows at least quadratically in m, $\binom{n-m}{m-\ell}$ for $\ell = 0$ dominates all the other summands as n (and thereby also m) approaches ∞.

Proof. Denote by s_ℓ the ℓ'th summand, i.e.,

$$s_\ell := \binom{m}{\ell}\binom{n-m}{m-\ell} K^\ell.$$

It then holds for all $\ell \in \{1, \dots, m-1\}$ that

$$\begin{aligned}
\frac{s_{\ell+1}}{s_\ell} &= \frac{\binom{m}{\ell+1}\binom{n-m}{m-\ell-1}K^{\ell+1}}{\binom{m}{\ell}\binom{n-m}{m-\ell}K^\ell} \\
&= \frac{\frac{m!}{(\ell+1)!(m-\ell-1)!}\frac{(n-m)!}{(m-\ell-1)!(n-2m+\ell+1)!}}{\frac{m!}{\ell!(m-\ell)!}\frac{(n-m)!}{(m-\ell)!(n-2m+\ell)!}}K \\
&= \frac{(m-\ell)^2}{(\ell+1)(n-2m+\ell+1)}K \\
&\leq \frac{m^2}{2(n-2m+2)}K.
\end{aligned}$$

Since, by assumption, $m = o(\sqrt{n})$, this converges to 0 as n goes to infinity. In particular, there exists a constant $c \in (0,1)$ such that for n sufficiently large it holds for all $\ell \in \{1, \dots, m-1\}$ that $\frac{s_{\ell+1}}{s_\ell} \leq c$. This implies that $s_\ell \leq s_1 c^{\ell-1}$, and hence also

$$\sum_{\ell=1}^{m} s_\ell \leq s_1 \sum_{\ell=1}^{m} c^{\ell-1} \leq s_1 \frac{1}{1-c},$$

where for the last inequality we used the explicit solution of a geometric sum. We now conclude the proof by explicitly bounding (A.10) as follows

$$\begin{aligned}
\frac{1}{\binom{n}{m}}\sum_{\ell=1}^{m}\binom{m}{\ell}\binom{n-m}{m-\ell}K^\ell &= \frac{1}{\binom{n}{m}}\sum_{\ell=1}^{m} s_\ell \\
&< \frac{1}{1-c}\frac{s_1}{\binom{n}{m}} \\
&= \frac{K}{1-c}\frac{m\binom{n-m}{m-1}}{\binom{n}{m}} \\
&= \frac{K}{1-c}\frac{m^2}{n}\frac{\binom{n-m}{m-1}}{\binom{n-1}{m-1}},
\end{aligned} \tag{A.11}$$

□

A.8.2. Proof of Theorem 2

Proof. We explicitly construct an example hypothesis test for which the worst case rate is achieved. We construct a target and observation density on $[0, \infty)$. First, for fixed $\alpha \in (0,1)$ and all $v \in \mathbb{N} \setminus \{0\}$ define

$$c_v := (1-\alpha)^{\frac{1}{v}} \quad \text{and} \quad p_v := 1 - (v+1)^{-\varepsilon},$$

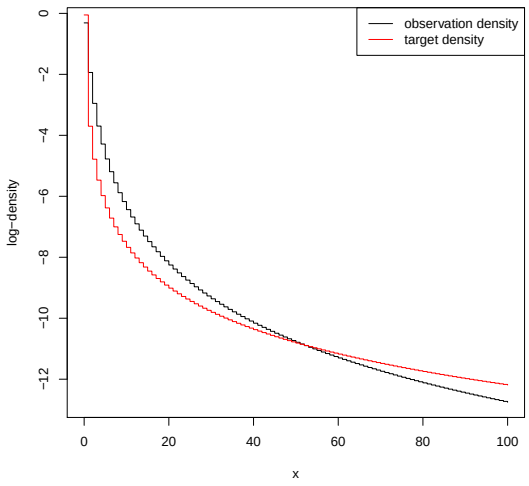


Figure A.4.: Visualization of densities in the proof of Theorem 2 with $\varepsilon = 1.9$. The tail of the target density eventually becomes larger than that of the observation density.

and $c_0 := 0$ and $p_0 := 0$, with $\varepsilon \in (0, \frac{\ell}{\ell-1})$ to be chosen below. Then, for all $v \in \mathbb{N}$ define

$$f_v := c_{v+1} - c_v \quad \text{and} \quad g_v := p_{v+1} - p_v.$$

Using these sequences, we define the following two densities:

(1) Target density (cdf is denoted by F): For all $x \in \mathbb{R}$, we define

$$f(x) := \sum_{v=0}^{\infty} \mathbb{1}_{\{v \leq x < v+1\}} f_v,$$

(2) Observation density (cdf is denoted by G): For all $x \in \mathbb{R}$, we define

$$g(x) := \sum_{v=0}^{\infty} \mathbb{1}_{\{v \leq x < v+1\}} g_v.$$

As $\lim_{v\to\infty} c_v = \lim_{v\to\infty} p_v = 1$, these functions are indeed densities. Moreover, one can verify that for all $m \in \mathbb{N}$, we have

$$F(m) = c_m \quad \text{and} \quad G(m) = p_m.$$

A visualization of the densities is given in Fig. A.4.

Finally, if we define, for all $x \geq 0$, $r(x) := \frac{f(x)}{g(x)}$ then we get

$$\begin{aligned}\mathbb{E}_g\left[r(X)^\ell\right] &= \int_0^\infty r(x)^\ell g(x)dx \\ &= \sum_{v=1}^\infty \left(\frac{f_v}{g_v}\right)^\ell g_v \\ &= \sum_{v=1}^\infty \frac{(c_{v+1}-c_v)^\ell}{(p_{v+1}-p_v)^{\ell-1}}.\end{aligned}$$

This series converges for all possible parameter choices $\varepsilon \in (0, \frac{\ell}{\ell-1})$ because

(a) $(c_{v+1} - c_v) \sim -\log(1-\alpha)v^{-2}$ as $v \to \infty$ and

(b) $(p_{v+1} - p_v) \sim \varepsilon v^{-(\varepsilon+1)}$ as $v \to \infty$.

(Indeed, both results follow from the mean value theorem as follows: First, for (a) applying the mean value theorem to $x \mapsto (1-\alpha)^{1/x}$ implies that for all $v \in \mathbb{N}$ there exists $\xi_v \in [v, v+1]$ such that

$$\frac{c_{v+1}-c_v}{v+1-v} = -\log(1-\alpha)\frac{(1-\alpha)^{1/\xi_v}}{\xi_v^2}.$$

We therefore get

$$\lim_{v\to\infty}(c_{v+1}-c_v)v^2 = -\log(1-\alpha)\lim_{v\to\infty}\frac{v^2(1-\alpha)^{1/\xi_v}}{\xi_v^2} = -\log(1-\alpha).$$

Similarly, for (b), we apply the mean value theorem to $x \mapsto 1-(x+1)^{-\varepsilon}$ to get that for all $v \in \mathbb{N}$ there is a $\xi_v \in [v, v+1]$ such that

$$\frac{p_{v+1}-p_v}{v+1-v} = \varepsilon(\xi_v+1)^{-(\varepsilon+1)}.$$

Again taking the limits we get

$$\lim_{v\to\infty}(p_{v+1}-p_v)v^{(\varepsilon+1)} = \varepsilon\lim_{v\to\infty}(\xi_v+1)^{-(\varepsilon+1)}v^{(\varepsilon+1)} = \varepsilon.$$

This completes the proofs of (a) and (b).)

Now, consider the null hypothesis

$$H_0 \coloneqq \{P_f\}.$$

For all $m \in \mathbb{N}$, we define the test $\phi_m : [0,\infty)^m \to \{0,1\}$ for all $x_1, \ldots, x_n \in [0,\infty)$ by

$$\phi_m(x_1,\ldots,x_m) \coloneqq \mathbb{1}_{\{\max(x_1,\ldots,x_m)\leq m\}}.$$

Then, it holds that

$$\begin{aligned}\mathbb{P}_f(\phi_m(X_1,\dots,X_m)=0) &= \mathbb{P}_f(\max(X_1,\dots,X_m)>m)\\ &= 1-\mathbb{P}_f(X_i\leq m)^m\\ &= 1-F(m)^m\\ &= 1-c_m^m\\ &= 1-[(1-\alpha)^{\frac{1}{m}}]^m\\ &= \alpha.\end{aligned}$$

Hence, ϕ_m achieves valid level in the target distribution f. Our goal is now to show that any resampling procedure for testing under distributional shifts with $m=n^q$ and $q>\frac{\ell-1}{\ell}$ cannot achieve asymptotic level. Let Ψ^m be the resampling scheme from the theorem that outputs a (not necessarily distinct) sample of size $m=n^q$. Then, it holds that

$$\begin{aligned}\mathbb{P}_g(\phi_m(\Psi^m(X_1,\dots,X_n))=1) &= \mathbb{P}_g(\max(\Psi^m(X_1,\dots,X_n))\leq m)\\ &\geq \mathbb{P}_g(\max(X_1,\dots,X_n)\leq m)\\ &= \mathbb{P}_g(X_i\leq m)^n\\ &= G(m)^n\\ &= p_m^n\\ &= (1-(m+1)^{-\varepsilon})^n\\ &= \exp(n\log(1-(m+1)^{-\varepsilon})).\end{aligned}$$

Taylor expanding $x\mapsto\log(x)$ in $x_0=1$ yields

$$\log(x)=\log(x_0)+\frac{1}{x_0}(x-x_0)+\frac{1}{2}\frac{-1}{\xi_x^2}(x-x_0)^2$$

for an $\xi_x\in(x,1)$. Plugging in $x=1-(m+1)^{-\varepsilon}$, we get

$$\log(1-(m+1)^{-\varepsilon})=-(m+1)^{-\varepsilon}-\frac{1}{2\xi_m^2}(m+1)^{-2\varepsilon},\tag{A.12}$$

where ξ_x is lower bounded by $(1-(m+1)^{-\varepsilon})$, and so $\xi_m\to 1$ and thus $\xi_m^2\to 1$ for $m\to\infty$. Next, observe that

$$n(m+1)^{-\varepsilon}\leq nm^{-\varepsilon}=n^{1-q\varepsilon}.$$

Now if we select $\frac{\ell-1}{\ell}>\varepsilon>1/q$ (this is always possible because $q>1/2$), we have that $q\varepsilon>1$, and it holds that $\lim_{n\to\infty}n^{1-q\varepsilon}=0$. For the same reason, we have $n(m+1)^{-2\varepsilon}\leq nm^{-\varepsilon}\to 0$ as $n\to\infty$. Combining this with the (A.12), we get

$$\lim_{n\to\infty}n\log(1-(m+1)^{-\varepsilon})=0$$

and thus

$$\lim_{n\to\infty} \mathbb{P}_g(\phi_m(\Psi^m(X_1,\dots,X_n)) = 1) \geq 1.$$

This completes the proof of Theorem 2. □

A.8.3. Proof of Theorem 3

Proof. The proof is similar to the proof of Theorem 1 but we will need to adjust for the estimation of the distributional shift factor. In particular, we will reprove the results in Lemma A.1 when using the estimator $\hat{r}_{n_1}$.

Fix any $P \in H_0$ and let $Q \in \tau^{-1}(\{P\})$. Denote by p and q their respective densities with respect to the dominating measure μ. We begin by recalling the details for the sample splitting procedure described in Algorithm A.1: $\mathbf{X}_n$ is split into two disjoint data sets $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ of sizes n_1, n_2, where $n_1 + n_2 = n$. The assumptions $n_1^a = \sqrt{n_2}$ and $m = o(\min(n^a, \sqrt{n}))$, ensure that $m = o(n_1^a)$ and $m = o(\sqrt{n_2})$.[2] We use $\mathbf{X}_{n_1}$ to fit an estimator $\hat{r}_{n_1}$ of r_q and then use $\mathbf{X}_{n_2}$ for the resampling. When taking expectations over $\mathbf{X}_{n_1}$ we use the notation $\mathbb{E}_{Q_1}$. Similarly, $\mathbb{E}_{Q_2}$ denotes an expectation over $\mathbf{X}_{n_2}$. We write $\mathbb{E}_Q$ when taking expectations with respect to the entire sample $\mathbf{X}_n$. Let $I_1 \coloneqq \{1,\dots,n_1\}$ and $I_2 \coloneqq \{n_1+1,\dots,n_1+n_2\}$ be the indices of $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ respectively.

Using the same argument as we used to derive (A.1) in the proof of Theorem 1, we get that

$$\mathbb{P}_Q(\varphi_m(\Psi^{\hat{r}_{n_1}}(\mathbf{X}_{n_2}, U) = 1)) = \mathbb{E}_Q\left[\frac{\frac{1}{\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_v})\right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}}{\frac{1}{\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_\ell})}\right], \tag{A.13}$$

where X_{i_ℓ} are observations from $\mathbf{X}_{n_2}$. As in the proof of Lemma A.1, we prove the convergence in probability of the numerator and denominator in (A.13) separately. Again, we do this in two steps: (A) We show that the means converge to the desired quantity and (B) we show that the variances converge to zero. First we show the following intermediate result.

Intermediate result: Let $\varepsilon(n_1) \coloneqq \sup_{x\in\mathcal{X}} \mathbb{E}_{Q_1}\left|\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{n_1^a} - 1\right|$ and consider a sequence $i_1,\dots,i_m$ from the indices of $\mathbf{X}_2$. Then, for n_1 sufficiently large and using

[2] When $n_1^a = \sqrt{n_2}$ and $n_1 + n_2 = n$, we have $n = n_1^{2a} + n_1$ and $n = n_2 + n_2^{1/(2a)}$. If $a > 1/2$, we have $m = o(\sqrt{n}) = o(\sqrt{n_1^{2a}}) = o(n_1^a)$, and $m = o(\sqrt{n}) = o(\sqrt{n_2})$. Similar arguments apply if $a < 1/2$.

Jensen's inequality, it holds Q_{n_2}-a.s. that

$$\begin{aligned}\left|\mathbb{E}_{Q_1}\left[\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right]-1\right| &\leq \mathbb{E}_{Q_1}\left[\left|\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}-1\right|\right]\\ &\leq \sup_{x_{i_1},\ldots,x_{i_m}\in\mathcal{X}}\mathbb{E}_{Q_1}\left[\left|\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(x_{i_\ell})}{r_q(x_{i_\ell})}-1\right|\right]\\ &\leq \sup_{x\in\mathcal{X}}\mathbb{E}_{Q_1}\left[\left|\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{m}-1\right|\right]\\ &\leq \sup_{x\in\mathcal{X}}\mathbb{E}_{Q_1}\left[\left|\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{n_1^a}-1\right|\right]\\ &= \varepsilon(n_1). \end{aligned} \tag{A.14}$$

The last inequality holds because by assumption, $m = o(n_1^a)$ when $n_1 \to \infty$, so for n_1 sufficiently large, $n_1^a > m$. For any $m, k \in \mathbb{N}$ we have $c^{m+k} \geq c^m \geq 1$ if $c > 1$ and $c^{m+k} \leq c^m \leq 1$ if $0 \leq c \leq 1$, and in either case it holds that $|c^{m+k} - 1| \geq |c^m - 1|$. Similarly, for $i_1, \ldots, i_m$ and $i'_1, \ldots, i'_m$ from I_2, we get Q_{n_2}-a.s. that

$$\begin{aligned}&\left|\mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right)\left(\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})}\right)\right]-1\right|\\ &\leq \mathbb{E}_{Q_1}\left[\left|\left(\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right)\left(\prod_{\ell=1}^{m}\frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})}\right)-1\right|\right]\\ &\leq \sup_{x\in\mathcal{X}}\mathbb{E}_{Q_1}\left[\left|\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{2m}-1\right|\right]\\ &\leq \sup_{x\in\mathcal{X}}\mathbb{E}_{Q}\left[\left|\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{n_1^a}-1\right|\right]\\ &= \varepsilon(n_1), \end{aligned} \tag{A.15}$$

using that for n_1 sufficiently large, $n_1^a > 2m$. This concludes the intermediate result.

Before showing parts A and B, we introduce some notation. Depending on whether we consider the numerator or denominator case, we define either $\delta_m := \mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}}$ or $\delta_m := 1$. Furthermore, we introduce for any function $r : \mathcal{X} \to (0, \infty)$ the following random variable

$$M(r) := \frac{1}{\frac{n_2!}{(n_2-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct from } I_2}}\left(\prod_{\ell=1}^{m} r(X_{i_\ell})\right)\delta_m.$$

Part A (means): Since $\hat{r}_{n_1}(x) = r_q(x)\frac{\hat{r}_{n_1}(x)}{r_q(x)}$, using the independence between $\mathbf{X}_{n_1}$

and $\mathbf{X}_{n_2}$ we get that

$$\begin{aligned}\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] &= \frac{1}{\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \mathbb{E}_Q\left[\left(\prod_{\ell=1}^m r_q(X_{i_\ell})\right)\left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right)\delta_m\right] \\ &= \frac{1}{\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \mathbb{E}_{Q_2}\left[\left(\prod_{\ell=1}^m r_q(X_{i_\ell})\right)\mathbb{E}_{Q_1}\left[\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right]\delta_m\right].\end{aligned} \tag{A.16}$$

We emphasize that the expectation $\mathbb{E}_{Q_1}$ only averages over the randomness in estimating $\hat{r}$, and does take expectations over X_{i_ℓ}. which is drawn from Q_{n_2}. Furthermore, using the intermediate result (A.14), we get the following upper bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] \leq \mathbb{E}_{Q_2}\left[M(r_q)\right](1+\varepsilon(n_1))$$

and lower bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] \geq \mathbb{E}_{Q_2}\left[M(r_q)\right](1-\varepsilon(n_1)).$$

Since $m = o(\sqrt{n_2})$, we can apply Lemma A.1 (a) and (b) to get that the means $\mathbb{E}_Q[M(\hat{r}_{n_1})]$ of the denominator and numerator converge to the desired values.

Part B (variances): Next, we show that both for the numerator and denominator the variance converges to zero. To this end, we expand the second moment as follows

$$\begin{aligned}\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] &= \mathbb{E}_Q\left[\frac{1}{\frac{n_2!}{(n_2-m)!}\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \sum_{\substack{i'_1,\dots,i'_m\\ \text{distinct from } I_2}} \left(\prod_{\ell=1}^m \hat{r}_{n_1}(X_{i_\ell})\right)\left(\prod_{\ell=1}^m \hat{r}_{n_1}(X_{i'_\ell})\right)\delta_m\delta'_m\right] \\ &= \mathbb{E}_{Q_2}\left[\frac{1}{\frac{n_2!}{(n_2-m)!}\frac{n_2!}{(n_2-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct from } I_2}} \sum_{\substack{i'_1,\dots,i'_m\\ \text{distinct from } I_2}} \left(\prod_{\ell=1}^m r(X_{i_\ell})\right)\left(\prod_{\ell=1}^m r(X_{i'_\ell})\right)\right. \\ &\qquad \left.\mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})}\right)\left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})}\right)\right]\delta_m\delta'_m\right].\end{aligned}$$

Here $\delta'_m := \mathbb{1}_{\{\varphi_m(X_{i'_1},\dots,X_{i'_m})=1\}}$. Using the intermediate result (A.15) we get the following upper bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] \leq \mathbb{E}_{Q_2}\left[M(r_q)^2\right](1+\varepsilon(n_1))$$

and lower bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] \geq \mathbb{E}_{Q_2}\left[M(r_q)^2\right](1-\varepsilon(n_1)).$$

In Lemma A.1 (c) and (d) we have shown that $\lim_{n\to\infty}\mathbb{V}_Q(M(r_q)) = 0$. Hence, combining these bounds on the second moment with the above bounds on the first moment shows that also $\lim_{n\to\infty}\mathbb{V}_Q(M(\hat{r}_{n_1})) = 0$. This completes the proof of Theorem 3. □

A.8.4. Proof of Theorem 4

Proof of Theorem 4. The first part of the proof follows that of Theorem 1. Let p and q denote the respective densities of P and Q with respect to the dominating measure μ. Recall that we call a sequence $(i_1, \dots, i_m)$ distinct if for all $\ell \neq \ell'$ we have $i_\ell \neq i_{\ell'}$. The resampling scheme $\Psi_{\texttt{DRPL}}$ samples from the space of distinct sequences $(i_1, \dots, i_m)$, where every sequence has probability $w_{(i_1,\dots,i_m)} \propto \prod_{\ell=1}^m r(X_{i_\ell})$. The normalization constant is the sum over the weights in the entire space of distinct sequences, that is,

$$w_{(i_1,\dots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}.$$

Thus, taking an expectation involving $\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$, amounts to evaluating φ_m in all distinct sequences $X_{i_1}, \dots, X_{i_m}$ and weighting with the probabilities $w_{(i_1,\dots,i_m)}$.

$$\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1) = \mathbb{E}_Q\left[\frac{\frac{1}{\frac{n!}{(n-m)!}} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \left(\prod_{\ell=1}^m r(X_{i_\ell})\right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}}{\frac{1}{\frac{n!}{(n-m)!}} \sum\limits_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}\right], \tag{A.17}$$

where we divide by the number of distinct sequences $\frac{n!}{(n-m)!}$ in both numerator and denominator.

Let $c(n, m)$ and $d(n, m)$ be the numerator and denominator terms of (A.17), i.e.,

$$c(n,m) \coloneqq \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \left(\prod_{\ell=1}^m r(X_{i_\ell})\right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}},$$

$$d(n,m) \coloneqq \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(j_1,\dots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell}).$$

Define for all $\delta > 0$ the set $A_\delta \coloneqq \{d(n, m) \geq 1 - \delta\}$. It holds for all $\delta \in (0, 1)$ that

$$\begin{aligned}
\mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] &= \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\leq \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\leq \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\right] + \mathbb{P}_Q\left(A_\delta^c\right) \\
&= \frac{1}{1-\delta}\mathbb{P}_P(\varphi_m(X_1,\dots,X_m) = 1) + \mathbb{P}_Q\left(A_\delta^c\right),
\end{aligned}$$

where we used that $\frac{c(n,m)}{d(n,m)} \leq 1$ and Lemma A.1 (a). Further, by applying Cantelli's inequality to $\mathbb{P}_Q(A_\delta^c)$, it follows that

$$\mathbb{P}_Q\left(A_\delta^c\right) \leq \frac{\mathrm{VAR}_Q(d(n,m))}{\mathrm{VAR}_Q(d(n,m)) + \delta^2}.$$

Finally, we can apply (A.7),

$$\begin{aligned} V(n,m) \coloneqq \mathrm{VAR}_Q(d(n,m)) &= \mathrm{VAR}\left(\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^{m} r(X_{i_\ell})\right)\right) \\ &= \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} (\mathbb{E}_Q\left[r(X_{i_1})^2\right]^\ell - 1), \end{aligned}$$

where we use that ζ_v (used in (A.7)) is given by

$$\begin{aligned} \zeta_v &= \mathbb{V}_Q\left(\mathbb{E}_Q\left[\prod_{\ell=1}^{m} r(X_{i_\ell}) \mid X_{i_1}, \ldots, X_{i_v}\right]\right) \\ &= \mathbb{V}_Q\left(\left(\prod_{\ell=1}^{v} r(X_{i_\ell})\right) \mathbb{E}_Q\left[\prod_{\ell=v+1}^{m} r(X_{i_\ell})\right]\right) \\ &= \mathbb{V}_Q\left(\prod_{\ell=1}^{v} r(X_{i_\ell})\right) \\ &= \mathbb{E}_Q[r(X_{i_1})^2]^v - 1. \end{aligned}$$

Plugging in this upper bound for $\mathbb{P}_Q(A_\delta^c)$ yields

$$\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1) = \frac{1}{1-\delta}\mathbb{P}_P(\varphi_m(X_1, \ldots, X_m) = 1) + \frac{V(n,m)}{V(n,m) + \delta^2}.$$

Since $\delta \in (0,1)$ was arbitrary, the theorem statement follows. □

A.8.5. Proof of Theorem 5

Proof. We adjust part of the proof of Theorem 1 to the uniform case. Again, let $c(n,m)$ and $d(n,m)$ be the numerator and denominator terms of (A.1), i.e.,

$$c(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\dots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) \right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}},$$

$$d(n,m) := \frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(j_1,\dots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{j_\ell}).$$

We want to show that $\limsup_{n\to\infty} \sup_{Q\in\tau^{-1}(H_0)} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] \leq \alpha_\varphi$. To see this, define for all $\delta > 0$ the set $A_\delta := \{|d(n,m) - 1| \leq \delta\}$, and take any $P \in H_0$ and $Q \in \tau^{-1}(\{P\})$. It holds for all $\delta \in (0,1)$ that

$$\begin{aligned}
\mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] &= \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\leq \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\mathbb{1}_{A_\delta}\right] + \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\mathbb{1}_{A_\delta^c}\right] \\
&\leq \mathbb{E}_Q\left[\frac{c(n,m)}{1-\delta}\right] + \mathbb{P}_Q\left(A_\delta^c\right) \\
&= \frac{1}{1-\delta}\mathbb{P}_P(\varphi_m(X_1,\dots,X_m)=1) + \mathbb{P}_Q\left(A_\delta^c\right),
\end{aligned}$$

where we used that $\frac{c(n,m)}{d(n,m)} \leq 1$ and that $\mathbb{E}_Q[c(n,m)] = \mathbb{P}_P(\varphi_m(X_1,\dots,X_m)=1)$, as shown in Lemma A.1 (a). Further, given the uniform bound on the weights, combining Chebyshev's inequality with Lemma A.1 (b) and (d) leads to $\lim_{n\to\infty} \sup_{Q\in\tau^{-1}(H_0)} \mathbb{P}_Q\left(A_\delta^c\right) = 0$. Hence, using that φ has uniform asymptotic level α_φ we have shown for all $\delta \in (0,1)$ that

$$\limsup_{n\to\infty} \sup_{Q\in\tau^{-1}(H_0)} \mathbb{E}_Q\left[\frac{c(n,m)}{d(n,m)}\right] \leq \frac{1}{1-\delta}\alpha_\varphi.$$

Using that $\delta \in (0,1)$ is arbitrary, completes the proof of Theorem 5. □

A.8.6. Proof of Corollary A.1

Proof. We have

$$\begin{aligned}
&\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U))=1) \\
&\quad= \mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U))=1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U) \text{ distinct})\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U) \text{ distinct}) \\
&\qquad+ \mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U))=1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U) \text{ not distinct})\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n,U) \text{ not distinct}).
\end{aligned}$$

This converges to the same limit as $\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)) = 1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)\,\text{distinct})$ (because $\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)\,\text{distinct})$ converges to 1, see Proposition A.1), which, as we argue in Appendix A.4, equals $\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1)$. The result then follows from Theorem 1. □

A.8.7. Proof of Proposition A.1

Proof. When sampling with replacement, $\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)$ contains non-distinct draws with positive probability (assuming, wlog, that m is not 1). Yet, we show that $\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)$ distinct) approaches 1 as $m \to \infty$. By assumption $p = \tau(q)$, so $p(x) \propto r(x)q(x)$. Let $\bar{r}$ be the normalized version of r satisfying, for all x, $p(x) = \bar{r}(x)q(x)$. The probability $w_{(i_1,\ldots,i_m)}$ of drawing a sequence $X_{i_1}, \ldots, X_{i_m}$ is defined by (10) as the product of weights r:

$$w_{(i_1,\ldots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} = \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})},$$

where the sum over $(j_1, \ldots, j_m)$ in the denominator is over all sequences of length m (including distinct and non-distinct sequences). The probability of drawing a non-distinct sequence equals the sum of the weights corresponding to all non-distinct sequences $w_{(i_1,\ldots,i_m)}$. Therefore,

$$\begin{aligned}
&\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)\,\text{not distinct}) \\
&= \mathbb{E}_Q\left[\sum_{\substack{(i_1,\ldots,i_m)\\ \text{not distinct}}} w_{(i_1,\ldots,i_m)}\right] \\
&= \mathbb{E}_Q\left[\sum_{\substack{(i_1,\ldots,i_m)\\ \text{not distinct}}} \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})}\right] \\
&= \mathbb{E}_Q\left[\frac{\sum_{\substack{(i_1,\ldots,i_m)\\ \text{not distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{(i_1,\ldots,i_m)} \prod_{\ell=1}^m \bar{r}(X_{i_\ell})}\right] \\
&= \mathbb{E}_Q\left[\frac{\frac{1}{n^m}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{not distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\frac{1}{n^m}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{not distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell}) + \frac{1}{n^m}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell})}\right].
\end{aligned} \tag{A.18}$$

Observe that this expectation is taken both over $\mathbf{X}_n$ and U. By Lemma A.3, the numerator of (A.18) (which equals the first term in the denominator) converges to 0 in L^1. The

second term in the denominator converges in probability to 1 by Lemma A.1 (this requires Assumption (A2), which is implied by Assumption (A3)); thus, the entire denominator converges to 1 in probability. By Slutsky's lemma, the entire fraction (inside the mean) converges to 0 in probability. Since the fraction is lower bounded by 0 and upper bounded by 1, convergence in probability implies convergence of the mean (see the proof of Theorem 1 for an argument for this), and it follows that $\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ not distinct}) \to 0$. □

Lemma A.3 (Non-distinct draws)**.** *Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ be distributions with densities p and q with respect to a dominating measure μ. Let $\bar{r} : \mathcal{X} \to (0, \infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) = \bar{r}(x)q(x)$. Then, under Assumptions (A1) and (A3) it holds that*

$$\lim_{n\to\infty} \mathbb{E}_Q \left[\frac{1}{n^m} \sum_{\substack{(i_1,\dots,i_m) \\ \textit{not distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) \right] = 0.$$

In particular, since the integrand is non-negative, this implies that

$$\frac{1}{n^m} \sum_{\substack{(i_1,\dots,i_m) \\ \textit{not distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) \xrightarrow{L^1} 0 \quad \textit{as } n \to \infty.$$

Proof. We first rewrite the sum using the number k of distinct draws, i.e., we consider cases, in which there are k distinct elements among $i_1, \dots, i_m$. The number k is at least 1 and, since not all draws are distinct, at most $m-1$. For fixed k, we then further sum over the numbers $r_1, \dots, r_k$ of occurrences of each index, i.e., j_ℓ appears r_ℓ times.

$$\sum_{\substack{(i_1,\dots,i_m) \\ \text{not distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) = \sum_{k=1}^{m-1} \sum_{\substack{(i_1,\dots,i_m) \text{ with } k \text{ distinct entries} \\ (\text{i.e. } j_\ell \in (i_1,\dots,i_m) \text{ appears } r_\ell > 0 \text{ times}, \\ r_1+\dots+r_k=m)}} \prod_{\ell=1}^{k} \bar{r}(X_{j_\ell})^{r_\ell}.$$

Using the independence across distinct observations, this implies that

$$\mathbb{E}_Q \left[\sum_{\substack{(i_1,\dots,i_m) \\ \text{not distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{i_\ell}) \right] = \sum_{k=1}^{m-1} \sum_{\substack{(i_1,\dots,i_m) \text{ with } k \text{ distinct entries} \\ (\text{i.e. } j_\ell \in (i_1,\dots,i_m) \text{ appears } r_\ell > 0 \text{ times}, \\ r_1+\dots+r_k=m)}} \prod_{\ell=1}^{k} \mathbb{E}_Q \left[\bar{r}(X_{j_\ell})^{r_\ell} \right]. \tag{A.19}$$

We now use the uniform bound on the weights given in Assumption (A3) and the fact that $\mathbb{E}_Q[\bar{r}(X_i)^t] = \int \bar{r}(x_i)q(x_i)\bar{r}(x_i)^{t-1}\mathrm{d}\mu(x_i) = \int p(x_i)\bar{r}(x_i)^{t-1}\mathrm{d}\mu(x_i) = \mathbb{E}_P[\bar{r}(X_i)^{t-1}]$ to get for all $i \in \{1, \dots, n\}$ and all $t \in \{1, \dots, m-1\}$ that

$$\mathbb{E}_Q[\bar{r}(X_i)^t] = \mathbb{E}_P[\bar{r}(X_i)^{t-1}] \leq L^{t-1}.$$

$$\mathbb{E}\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell})\right] \leq \sum_{k=1}^{m-1}\sum_{\substack{(i_1,\dots,i_m) \text{ with } k \text{ distinct entries}\\ (\text{i.e. } j_\ell\in(i_1,\dots,i_m) \text{ appears } r_\ell>0 \text{ times},\\ r_1+\dots+r_k=m)}} L^{m-k}$$

$$= \sum_{k=1}^{m-1}\binom{n}{k}\pi(m,k)L^{m-k}$$

$$\leq \sum_{k=1}^{m-1}\binom{n}{k}\tilde{\pi}(m,k)L^{m-k}, \tag{A.20}$$

where $\pi(m,k)$ is the number of words of length m using k letters such that each letter is used at least once and

$$\tilde{\pi}(m,k) \coloneqq k^{m-k}\frac{m!}{(m-k)!}.$$

The last inequality holds because we have $\pi(m,k) \leq \tilde{\pi}(m,k)$: Consider constructing a word of length m by first distributing one of each of the k letters (ensuring that each letter is used at least once) among the m positions, which can be done in $m!/(m-k)!$ ways. For the remaining $m-k$ positions pick any combination of letters, which can be done in k^{m-k} ways. In total, this two-step procedure has $\tilde{\pi}(m,k)$ possible outcomes. This enumeration contains all words of length m using k letters such that each is used at least once, so $\pi(m,k) \leq \tilde{\pi}(m,k)$. We do not have equality, because $\tilde{\pi}$ counts some words several times, but with different intermediate steps. For example if $k=2$ and $m=3$, $\tilde{\pi}$ counts $(a,\_,b)+(\_,a,\_)$ and $(\_,a,b)+(a,\_,\_)$ as two distinct words, although they both yield (a,a,b); indeed $\pi(3,2)=6$ and $\tilde{\pi}(3,2)=12$.

Then, with $s_k \coloneqq \binom{n}{k}\tilde{\pi}(m,k)L^{m-k}$ it holds that

$$\begin{aligned}
\frac{s_{k+1}}{s_k} &= \frac{\binom{n}{k+1}\tilde{\pi}(m,k+1)L^{m-k-1}}{\binom{n}{k}\tilde{\pi}(m,k)L^{m-k}} \\
&= \frac{1}{L}\frac{n-k}{k+1}\frac{m-k}{1}\frac{(k+1)^{m-k-1}}{k^{m-k}} \\
&= \frac{1}{L}\frac{n-k}{k+1}\frac{m-k}{k+1}\left(\frac{k+1}{k}\right)^{m-k} \\
&\geq \frac{1}{L}\frac{n-m+1}{m^2} \\
&=: c,
\end{aligned}$$

where the inequality follows by using $k \leq m-1$ and $(k+1)/k \geq 1$. By Assumption (A1) (i.e., $m = o(\sqrt{n})$) it holds for n sufficiently large that $c > 1$. Iterating this inequality, we get (again for n sufficiently large) that $s_k \leq c^{-(m-1-k)}s_{m-1}$, which we can plug into

(A.20) to get

$$\mathbb{E}\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell})\right] \leq \sum_{k=1}^{m-1} s_k \leq s_{m-1}\sum_{k=1}^{m-1} c^{-(m-1-k)} = s_{m-1}\sum_{k=0}^{m-2} c^{-k} \leq s_{m-1}\frac{1}{1-\frac{1}{c}}. \tag{A.21}$$

In the last inequality, we use the trivial bound $\sum_{k=0}^{m-2} c^{-k} < \sum_{k=0}^{\infty} c^{-k}$ and $0 < c^{-1} < 1$. Finally, observe that

$$\begin{aligned} n^{-m}s_{m-1} &= n^{-m}\binom{n}{m-1}\tilde{\pi}(m,m-1)L \\ &= n^{-m}\frac{n!}{(n-(m-1))!(m-1)!}(m-1)^{m-(m-1)}\frac{m!}{(m-(m-1))!}L \\ &= n^{-m}\frac{n!}{(n-m)!}\frac{m(m-1)}{(n-m+1)}L \\ &= \underbrace{g(n,m)}_{:=n^{-m}\frac{n!}{(n-m)!}}\frac{m(m-1)}{(n-m+1)}L, \end{aligned}$$

which by Lemma A.4 converges to zero (by the assumption $m = o(\sqrt{n})$). Therefore, we have that

$$\mathbb{E}\left[n^{-m}\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar{r}(X_{i_\ell})\right] \leq n^{-m}s_{m-1}\frac{1}{1-\frac{1}{c}} \to 0,$$

which completes the proof of Lemma A.3. □

Lemma A.4. *Define for all $n, m \in \mathbb{N}$ the function*

$$g(n,m) \coloneqq \frac{n!}{(n-m)!}n^{-m}.$$

Then, it holds that

$$\lim_{n\to\infty} g(n,n^q) = \begin{cases} 0 & \text{if } q \in (\frac{1}{2},1) \\ \exp(-\frac{1}{2}) & \text{if } q = \frac{1}{2} \\ 1 & \text{if } q \in [0,\frac{1}{2}). \end{cases}$$

Proof. First, apply the Stirling approximation to get for n sufficiently large that

$$\begin{aligned} g(n,m) &\sim n^{n+\frac{1}{2}} \cdot e^{-n} \cdot (n-m)^{m-n-\frac{1}{2}} \cdot e^{n-m} \cdot n^{-m} \\ &= n^{n-m+\frac{1}{2}} \cdot (n-m)^{m-n-\frac{1}{2}} \cdot e^{-m} \\ &= \exp\{(n-m+\tfrac{1}{2})\log(n) + (m-n-\tfrac{1}{2})\log(n-m) - m\}. \end{aligned}$$

Next, we look at cases where $m = n^q$ for some $q \in [0,1)$. The above expression can then be simplified further as

$$\begin{aligned} g(n,n^q) &\sim \exp\{(n-n^q+\tfrac{1}{2})\log(n) + (n^q-n-\tfrac{1}{2})\log(n-n^q) - n^q\} \\ &= \exp\{(n-n^q+\tfrac{1}{2})\log(n) + (n^q-n-\tfrac{1}{2})[\log(n) + \log(1-n^{q-1})] - n^q\} \\ &= \exp\{(n^q-n-\tfrac{1}{2})\log(1-n^{q-1}) - n^q\}. \end{aligned}$$

Finally, since $n^{q-1} \to 0$ as n goes to infinity we can use the following Taylor expansion

$$\log(1-n^{q-1}) = -n^{q-1} - \tfrac{1}{2}n^{2(q-1)} + O(n^{3(q-1)}),$$

which results in

$$\begin{aligned} g(n,n^q) &\sim \exp\{(n^q-n-\tfrac{1}{2})\log(1-n^{q-1}) - n^q\} \\ &= \exp\{(n^q-n-\tfrac{1}{2})(-n^{q-1} - \tfrac{1}{2}n^{2(q-1)} + O(n^{3(q-1)})) - n^q\} \\ &= \exp\{-n^{2q-1} - \tfrac{1}{2}n^{3q-2} + n^q + \tfrac{1}{2}n^{3q-2} + \tfrac{1}{2}n^{q-1} + \tfrac{1}{4}n^{2q-2} + O(n^{2q-1}) - n^q\} \\ &= \exp\{-\tfrac{1}{2}n^{2q-1} + O(n^{3q-2})\}. \end{aligned}$$

From this we see that

$$\lim_{n\to\infty} g(n,n^q) = \begin{cases} 0 & \text{if } q \in (\frac{1}{2},1) \\ \exp(-\frac{1}{2}) & \text{if } q = \frac{1}{2} \\ 1 & \text{if } q \in [0,\frac{1}{2}). \end{cases}$$

This completes the proof of Lemma A.4. □

A.8.8. Proof of Corollary 1

As discussed in Section 2.3, the proposed procedure in Section 4.1 can also be used to construct a test for a hypothesis $H_0^{\mathcal{Q}}$ in the observed domain, satisfying the same theoretical guarantees.

Corollary A.2 (Pointwise level in the observed domain - detailed version)**.** *Consider hypotheses $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$ and $H_0^{\mathcal{P}} \subseteq \mathcal{P}$ in the observational and in the target domain, respectively. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which there exist a known map $r : \mathcal{X} \to (0,\infty)$ and a set A satisfying for all $q \in \mathcal{Q}$ and all $x \in \mathcal{Z}$ that $\tau(q)(x) \propto r(x^A)q(x)$, see (5). Assume $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$. Let φ_k be a sequence of tests for $H_0^{\mathcal{P}}$ with pointwise asymptotic level α_φ. Let $m = m(n)$ be a resampling size and let ψ_n^r be the DRPL-based*

resampling test defined by $\psi_n^r(\mathbf{X}_n, U) \coloneqq \varphi_m(\Psi_{DRPL}^{r,m}(\mathbf{X}_n, U))$*, see Algorithm 1. Then, if* m *satisfies Assumption (A1) and all* $Q \in H_0^{\mathcal{Q}}$ *satisfy Assumption (A2), it holds that*

$$\sup_{Q \in H_0^{\mathcal{Q}}} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha_\varphi,$$

i.e., ψ_n^r *satisfies pointwise asymptotic level* α *for the hypothesis* $H_0^{\mathcal{Q}}$*.*

Clearly, the condition $H_0^{\mathcal{Q}} \subseteq \tau(H_0^{\mathcal{P}})$ is satisfied when $H_0^{\mathcal{P}} = \tau(H_0^{\mathcal{Q}})$. This is the case for the conditional independence test described in Section 3.1, for example.

Proof. We have

$$\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}} \;\Rightarrow\; H_0^{\mathcal{Q}} \subseteq \tau^{-1}(H_0^{\mathcal{P}})$$

and therefore

$$\sup_{Q \in H_0^{\mathcal{Q}}} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n^r = 1) \leq \sup_{Q \in \tau^{-1}(H_0^{\mathcal{P}})} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n^r = 1).$$

Since Assumption (A2) is satisfied for all $Q \in H_0^{\mathcal{Q}}$, the statement follows from Theorem 1. This completes the proof of Corollary 1. □

A.8.9. Proof of Proposition 1

Proof. We analyze the output of Algorithm 2. For each $i \in \{1, \ldots, n\}$, we discard X_i if $U_i > \frac{r(X_i)}{M}$, where U_i is uniform on $(0, 1)$. The probability of the event E_i that X_i is not discarded equals

$$q(E_i) = \int q(E_i|x_i)q(x_i)\mathrm{d}x_i = \int \tfrac{r(x_i)}{M} q(x_i)\mathrm{d}x_i = \int \tfrac{c}{M}\tfrac{p(x_i)}{q(x_i)} q(x_i)\mathrm{d}x_i = \tfrac{c}{M},$$

where c is a constant such that $r(x)q(x) = cp(x)$. and the conditional density of X_i given that the sample is not discarded, $q(x_i|E_i)$, is given by

$$q(x_i|E_i) = \frac{q(x_i)}{q(E_i)} q(E_i|x_i) = q(x_i)\frac{M}{c}\frac{c}{M}\frac{p(x_i)}{q(x_i)} = p(x_i)$$

If $X_{i_1}, \ldots, X_{i_m}$ are the points that are not discarded, this means that $(X_{i_1}, \ldots, X_{i_m})$ is distributed as if it was m i.i.d. draws from P^* (where m is random). In particular, by the assumption that the probability that for all $k \in \mathbb{Z}$: $\mathbb{P}_P(\varphi_k(\mathbf{Z}_k) = 1) = \alpha_\varphi$ for k i.i.d. samples $\mathbf{Z}_k$ from $\mathbb{P}_P$, it follows that $\mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) = \mathbb{P}_Q(\varphi(X_{i_1}, \ldots, X_{i_m})) = \alpha_\varphi$. □

A.9. Analyzing Assumption (A2) in a linear Gaussian model

In this section, we show conditions for assumption Assumption (A2) to be satisfied when we consider the shift that changes a Gaussian conditional into a marginal, independent Gaussian target distribution.

Proposition A.2. *Consider a linear Gaussian setting where $Y = X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $X \sim \mathcal{N}(0, \sigma_X^2)$, with $\sigma_\varepsilon, \sigma_X$ known. Assume that we are interested in the distributional shift that replaces the conditional $q(y|x)$ (an $\mathcal{N}(x, \sigma_\varepsilon^2)$-density, evaluated at y) with an independent $\mathcal{N}(0, \sigma^2)$ target distribution $p(y)$. Formally, define the shift factor r for all $x, y \in \mathbb{R}$ as*

$$r(x, y) = \frac{p(y)}{q(y|x)}.$$

Then Assumption (A2) is satisfied for Q if and only if

$$\sigma^2 < 2(\sigma_\varepsilon^2 - \sigma_X^2).$$

Proof. We begin by directly expanding the second moment of the factor r under the observational distribution Q as follows,

$$\begin{aligned}
\mathbb{E}_Q\left[r(X,Y)^2\right] &= \mathbb{E}_Q\left[\left(\frac{p(Y)}{q(Y|X)}\right)^2\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\left(\frac{Y-X}{\sigma_\varepsilon}\right)^2 - \left(\frac{Y}{\sigma}\right)^2\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\left(\frac{\varepsilon}{\sigma_\varepsilon}\right)^2 - \left(\frac{X+\varepsilon}{\sigma}\right)^2\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\varepsilon^2\left(\frac{1}{\sigma_\varepsilon^2} - \frac{1}{\sigma^2}\right) - \frac{X^2}{\sigma^2} - \frac{2X\varepsilon}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{X^2}{\sigma^2} - \frac{2\varepsilon}{\sigma^2}X - \frac{\varepsilon^2}{\sigma^2} + \frac{\varepsilon^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2} - \frac{\varepsilon^2}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W + \frac{\varepsilon^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2} - \frac{\varepsilon^2}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right], \qquad \text{(A.22)}
\end{aligned}$$

where $W \coloneqq (X/\sigma_X + \varepsilon/\sigma_X)^2$. Next, observe that, conditioned on ε, W has a non-central $\chi^2_{(1)}$-distribution with non-centrality parameter ε^2/σ_X^2. The moment generating function of W is given by $M_W(t) = (1-2t)^{-1/2}\exp\left(\frac{\varepsilon^2}{\sigma_X^2}\frac{t}{1-2t}\right)$ for all $t < 1/2$. Hence, continuing

the computation in (A.22) and by conditioning on ε, we get

$$\begin{aligned}
\mathbb{E}_Q\left[r(X,Y)^2\right] &= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W\right)\Big|\varepsilon\right]\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[M_W\left(-\frac{\sigma_X^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}\mathbb{E}_Q\left[\exp\left(\frac{\varepsilon^2}{\sigma_X^2}\frac{-\sigma_X^2}{\sigma^2}\frac{1}{1+2\frac{\sigma_X^2}{\sigma^2}}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}\mathbb{E}_Q\left[\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\left(1-\frac{\sigma_\varepsilon^2}{\sigma^2+2\sigma_X^2}\right)\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}M_S\left(1-\frac{\sigma_\varepsilon^2}{\sigma^2+2\sigma_X^2}\right),
\end{aligned}$$

where $S \coloneqq (\varepsilon/\sigma_\varepsilon)^2$ and M_S is the moment generating function of a (central) $\chi^2_{(1)}$ distribution. $M_S(t)$ is finite if and only if $t < 1/2$, corresponding to $1 - \frac{\sigma_\varepsilon^2}{\sigma^2+2\sigma_X^2} < 1/2$ which is equivalent to $\sigma^2 < 2(\sigma_\varepsilon^2 - \sigma_X^2)$. □

B. Appendix to Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

Appendix

This appendix is structured as follows:

- In Appendix B.1, we provide details on the synthetic lab testing example, including how we generate the loss landscape in Fig. 1 (right).
- In Appendix B.2, we provide a "user's guide" to defining and interpreting parametric shifts, including worked examples for many common conditional distributions, as well as guidance on how to define and interpret the shift functions $s(Z; \delta)$.
- In Appendix B.3, we provide additional details on the worst-case optimization problem, as well as comparisons of the reweighting-based approach to the Taylor approximation approach. We also demonstrate that the quadratic approximation is exact, for particularly simple structural causal models.
- In Appendix B.4, we compare our approach to that of worst-case conditional subpopulation shifts, in the context of a simpler laboratory testing example where we can explicitly compute the worst-case conditional subpopulations. Here, we demonstrate that our approach can capture more realistic intuition regarding which shifts are plausible in practice.
- In Appendix B.5, we give additional experimental details, as well as illustrative samples from the generative model, for the CelebA experiment described in Section 4.
- In Appendix B.6, we give proofs for all the results in the main paper.

B.1. Details of Fig. 1

In Fig. 1 (right), we consider the following, artificial, generative model, which resembles the setup in Section 4.1, but with the addition of age as a continuous variable.

$$\begin{aligned}
\text{Age} &\sim \mathcal{N}(0, 0.5^2) \\
\mathbb{P}(\text{Disease} = 1|\text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\
\mathbb{P}(\text{Order} = 1|\text{Disease},\ \text{Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\
\text{Test Result}|\text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1)
\end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero. In Fig. 1 (right), we consider a simple predictive model: If lab tests are not available (Order $= 0$), this model predicts disease based on an unregularized logistic regression model, which uses age to predict disease. If a lab test is available, then it uses both age and the lab test for prediction. This model is trained on 100,000 samples from the training distribution. To construct the loss landscape shown in Fig. 1 (right), we first observe that

$$\mathbb{P}(O = 1|\text{Disease}, \text{Age}) = \text{sigmoid}(\eta(\text{Disease}, \text{Age})),$$

where that $\eta(\text{Disease}, \text{Age}) = 2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1$. We construct shifts using the shift function $s(\text{Disease}, \text{Age}; \delta) = \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}$, and for a grid of values for $(\delta_0, \delta_1) \in [-5, 5]^2$ we consider perturbed distributions with a different conditional distribution of testing,

$$\mathbb{P}_\delta(O = 1|\text{Disease},\ \text{Age}) = \text{sigmoid}\left(\eta(\text{Disease}, \text{Age}) + \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}\right),$$

but where all other parts of the generative model are fixed. For each value of $(\delta_0, \delta_1) \in [-5, 5]^2$, we draw 10,000 samples from the corresponding distribution, and compute the negative log-likelihood of the original predictive model under this new distribution. The resulting surface is plotted in Fig. 1 (right).

B.2. A user's guide to defining parametric shifts

In this section, we discuss practical considerations in designing parametric shift functions for different distributions.

- In Appendix B.2.1, we give examples of conditional exponential families, illustrative shift functions, and how to interpret them.
- In Appendix B.2.2, we formalize the idea that one can choose shift functions which depend on additional variables, other than the causal parents of a variable W_i.
- In Appendix B.2.3 we give guidance on how to define shift functions when the parameters $\eta(Z)$ are constrained to lie in a particular domain, which is relevant for considering shifts such as changing the variance of a conditional Gaussian.

Table B.1.: Examples of conditional exponential family distributions.

Distribution	Parameter space	Sufficient statistic	Inverse parameter map
Binary(p)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$p(W = 1\|Z) = \text{sigmoid}(\eta(Z))$
Categorical($p_1, \ldots, p_k$)	$\eta(Z) \in \mathbb{R}^k$	$[T(W)]_i = \mathbf{1}\{W = i\}$	$\mathbb{P}(W = i\|Z) = [\text{softmax}(\eta(Z))]_i$
Poisson(λ)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$\lambda = \exp(\eta(Z))$
Gaussian(μ, σ^2)	$\eta(Z)_1 \in \mathbb{R}, \eta(Z)_2 < 0$	$T(W) = (W, W^2)$	$\mu(Z) = -\frac{\eta(Z)_1}{2\eta(Z)_2}, \sigma^2(Z) = -\frac{1}{2\eta(Z)_2}$
Gamma(α, β)	$\eta(Z)_1 > -1, \eta(Z)_2 < 0$	$T(W) = (\log W, W)$	$\alpha(Z) = \eta(Z)_1 + 1, \beta(Z) = -\eta(Z)_2$

B.2.1. Conditional exponential family models and interpretations of shifts

In this section, we give examples of exponential families and their sufficient statistics, and discuss design considerations in specifying the shift function $s(Z; \delta)$. Here, we restrict attention to shifts in a single variable, for ease of notation. In Table B.1 we give examples of conditional exponential families, along with their typical parameterizations. In the examples below, we review how shift functions $s(Z; \delta)$ impact these parameters, and how they can also be interpreted on the scale of more commonly considered parameters (e.g., conditional means and variances).

Example B.1 (Log-odds shift in a binary variable)**.** *Consider the distribution of a binary variable W conditioned on variables Z. Without loss of generality, we can write that*

$$\mathbb{P}(W = 1|Z) = \sigma(\eta(Z))$$

where σ is the sigmoid function, and $\eta(Z)$ is an arbitrary measurable function of Z, taking on values in the extended real line $\eta(Z) \in \mathbb{R} \cup \{-\infty, +\infty\}$. This can be written in canonical form as

$$\mathbb{P}(W|Z) = \exp\left\{\eta(Z) \cdot W - \log(1 + \exp^{\eta(Z)})\right\}$$

where $\eta(Z)$ is the canonical parameter (the log-odds ratio), $T(W) = W$ is the sufficient statistic, and $h(\theta) = \log(1+\exp^{\eta(Z)})$ is the normalizing constant. We can consider shifts $\eta_\delta(Z) \coloneqq \eta(Z) + \delta$, yielding the new conditional distribution

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta),$$

which is well-defined for any $\delta \in \mathbb{R}$.

Here, we note that these shifts occur on the "natural" parameter scale $\eta(Z)$ (e.g., the log-odds), which at first glance may seem difficult to interpret: Why should we care about changes on the log-odds scale, instead of on the original probability scale? In addition to mathematical convenience, we argue that in some settings, working with natural parameters is advantageous for retaining a common scale across across multiple variables.

For instance, consider shifts in the two independent variables W_1 and W_2, where $V_i \sim \text{Bernoulli}(p_i)$, with $p_1 = 10^{-4}$ and $p_2 = 0.6$. Suppose we wished to consider an

additive shift on the probability scale, e.g., $p'_1 = p_1 + 0.1, p'_2 = p_2 + 0.1$. Setting aside the inconvenience that we need to ensure $p'_1, p'_2 \in [0, 1]$, we argue that these shifts are not truly of a comparable scale. In particular, this shift in p_1 may seem implausible in magnitude, while the same shift in p_2 seems more reasonable. On the other hand, an additive shift in the log-odds captures some aspect of this idea.

Of course, there is some flexibility to incorporate prior expectations of shifts in absolute probabilities. For instance, in binary variable with no causal parents, we can always construct a one-to-one map of δ to a change in the marginal probability. For conditional shifts, we can similarly construct a one-to-one map between the value of δ in a shift $s(Z; \delta) = \delta$ and the resulting marginal probability of W_i, as formalized below.

Proposition B.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R} \cup \{+\infty, -\infty\}$. Let $p_+ \coloneqq \mathbb{P}(\eta(Z) = +\infty)$, $p_- \coloneqq \mathbb{P}(\eta(Z) = -\infty)$, and assume that $p_+ + p_- < 1$. Then, the marginal probability

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ whose range is $(p_+, 1 - p_-)$,

Proposition B.1 states that, for any achievable marginal probability $p_\delta = \mathbb{P}_\delta(W = 1)$, there exists a unique value of δ that achieves this probability. Because this relationship is strictly monotonic, we can hope to efficiently find such a value by e.g., binary search. In the laboratory testing example of Example 1, this would allow us to specify a plausible strength for the conditional shift δ in terms of an impact on the overall testing rate, e.g., modelling a scenario where the testing rate decreases from 20% to 15%.

Similar to the binary case, we can (if desired) directly parameterize shifts in terms of the conditional mean of a Gaussian distribution, as illustrated in Example B.2, which operates on the scale of $\mu(Z)$ alone.

Example B.2 (Mean shift in a conditional Gaussian)**.** *Consider the distribution of a multi-variate Gaussian variable W conditioned on a binary variable Z, where we write*

$$p(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z), \Sigma(z))$$

where $\mathcal{N}(w; \mu(z), \Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z; \delta) = [\Sigma(Z)^{-1}\delta, 0]$, such that

$$p_\delta(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z) + \delta, \Sigma(z)).$$

However, shifts of the same magnitude in the conditional mean may not be comparable.

Suppose that

$$\mathbb{P}(W|Z=0) \overset{(\mathrm{d})}{=} \mathcal{N}(0,1) \qquad \text{and} \qquad \mathbb{P}(W|Z=1) \overset{(\mathrm{d})}{=} \mathcal{N}(0,0.001),$$

such that $\delta = 1$ in Example B.2 corresponds to

$$\mathbb{P}_{\delta=1}(W|Z=0) \overset{(\mathrm{d})}{=} \mathcal{N}(1,1) \qquad \text{and} \qquad \mathbb{P}_{\delta=1}(W|Z=1) \overset{(\mathrm{d})}{=} \mathcal{N}(1,0.001).$$

While it may seem plausible that the mean of $W|Z=0$ can increase by 1, it may seem unrealistic for $W|Z=1$. Here, it may be more reasonable to consider a different parameterization of $s(Z;\delta)$, where the impact of the shift in a direction is proportional to the variance in that direction; we discuss this in the next example.

Example B.3 (Variance-scaled mean shift in a conditional Gaussian)**.** *Consider the distribution of a multi-variate Gaussian variable W conditioned on variables Z, where we write*

$$p(w|z) \overset{(d)}{=} \mathcal{N}(w;\mu(z),\Sigma(z))$$

where $\mathcal{N}(w;\mu(z),\Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z;\delta) = [\delta, 0]$, such that

$$p_\delta(w|z) \overset{(d)}{=} \mathcal{N}(w;\mu(z)+\delta^\top\Sigma(Z),\Sigma(z)).$$

In Example B.3, the parameter δ has a different interpretation, as a variance-scaled mean-shift. If W is one-dimensional, we can see that this becomes

$$p_\delta(w|z) \overset{(\mathrm{d})}{=} \mathcal{N}(w;\mu(z)+\delta\sigma^2(Z),\sigma^2(z)).$$

As we demonstrate in Appendix B.3.2, this particular example of a parameterization has other benefits: For instance, for estimation of shift gradients and Hessians at $\delta = 0$ can be done without knowledge of $\Sigma(Z)$.

B.2.2. Adding causal edges to the graph

In Section 2, we consider the case where the shift function $s(Z;\delta)$ alters a conditional $\mathbb{P}(W|Z)$ by a shift function $s(Z;\delta)$. We now discuss shift functions that use a larger set Z'. In particular, we consider the setting where Z represents the parents in a graph $\mathcal{G}$ (that is, $Z := \mathrm{PA}_\mathcal{G}(W)$), and consider shift functions that correspond to adding additional parents in that causal graph. Our definitions and results immediately extend to measuring the impact of shifts that **add edges** to the graph, in the form of shift functions that depend on non-descendants of W.

Building intuition with a simple example: To build intuition, consider the causal graph given in Fig. B.1. We consider a shift in X_2, with a shift function which depends

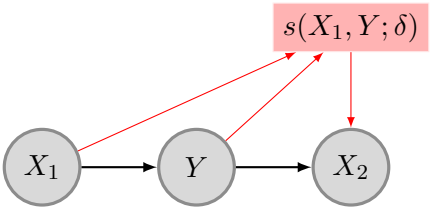


Figure B.1.: Illustrative example of an intervention $s(X_1, Y; \delta)$, and modified causal graph, which creates a dependence between X_1 and X_2 that bypasses Y.

not only on the causal parent Y, but also on X_1. Suppose that the distribution $\mathbb{P}(X_2|Y)$ is a conditional exponential family, given by

$$\mathbb{P}(X_2|Y) = g(X_2)\exp(\eta(Y)^\top T(X_2) - h(\eta(Y))).$$

Using that $X_2 \perp\!\!\!\perp X_1|Y$, we have $\mathbb{P}(X_2|Y) = \mathbb{P}(X_2|Y, X_2)$, and the joint probability factorizes as

$$\mathbb{P}(X_1, X_2, Y) = \mathbb{P}(X_2|Y)\mathbb{P}(Y|X_1)\mathbb{P}(X_1) = \mathbb{P}(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1).$$

This enables us to consider $Z = (Y, X_1)$ as the conditioning set in the context of Assumption 1. This is useful, because it allows us to consider shift functions that depend on Z, which includes X_1 in addition to Y. The δ-perturbation of this conditional distribution under the shift function $s(Y, X_1; \delta)$ is given by

$$\mathbb{P}_\delta(X_2|Y, X_1) = g(X_2)\exp\Big(\{\eta(Y) + s(Y, X_1; \delta)\}^\top T(X_2) - h\big(\eta(Y) + s(Y, X_1; \delta)\big)\Big),$$

and we can observe that under both graphs, the distribution factorizes in the same fashion, where

$$\mathbb{P}_\delta(X_1, X_2, Y) = \mathbb{P}_\delta(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1),$$

keeping the same convention that $s(Y, X_1; \delta = 0) = 0$, such that $\mathbb{P}_0 = \mathbb{P}$. This is one example of how our results can be applied with shift functions that effectively add edges to the causal graph. Of course, not all edges are permitted, so we give a more general treatment below.

General guidelines for adding edges: Allowing for the use of non-causal parents in the shift functions is straightforward, and can be done safely as follows, without violating Assumption 1: Given knowledge of the directed acyclic graph $\mathcal{G}$ which generates the observed distribution $\mathbb{P}$, we can **add** edges to the graph, as long as they do not create cycles.

Formally, let $\mathcal{G} = (\mathbf{V}, E)$ denote the causal DAG which generates the distribution $\mathbb{P}$, where $\mathbf{V}$ denotes variables and E denotes the set of edges, where we denote a directed edge by $e = (V_i, V_j)$, going from V_i to V_j. Let $\mathcal{G}' = (\mathbf{V}', E')$ denote another DAG (of our creation) with the constraint that we can only add edges, and that the graph must

remain acyclic, such that $E' \supseteq E$, and $\mathbf{V}' = \mathbf{V}$.

For any variable $W_i \in \mathbf{V}$, this implies that $\mathrm{PA}_{\mathcal{G}'}(W_i) \supseteq \mathrm{PA}_{\mathcal{G}}(W_i)$. Moreover, any new causal parent V_i of W_i in $\mathcal{G}'$ must have been a non-descendant of W_i in the original graph, as otherwise the graph $\mathcal{G}'$ would have a cycle from $W_i \to V_i \to W_i$. For ease of notation, let $N(W_i) \coloneqq \mathrm{PA}_{\mathcal{G}'}(W_i) \setminus \mathrm{PA}_{\mathcal{G}}(W_i)$ denote the set of new causal parents of W_i in $\mathcal{G}'$. For any variable W_i such that $N(W_i) \neq \varnothing$, we can write that

$$W_i \perp\!\!\!\perp_{\mathcal{G}} N(W_i) | \, \mathrm{PA}_{\mathcal{G}}(W_i) \tag{B.1}$$

by the rules of d-separation [Pearl, 2009]. As in Assumption 1, we use $\mathbf{W} = \{W_1, \ldots, W_m\}$ to denote the set of variables to be intervened upon, and accordingly will assume that in the causal graph $\mathcal{G}'$, we have not added new parents to any other variables, i.e., $N(V_i) = \varnothing$ for any $V_i \subsetneq \mathbf{W}$.

By (B.1), we can write that the distribution $\mathbb{P}$ factorizes as

$$\mathbb{P}(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i)) \right) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i | \, \mathrm{PA}_{\mathcal{G}}(V_i))$$

because $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i)) = \mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i)$, and if $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i))$ is a conditional exponential family satisfying Definition 2, then $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i))$ also satisfies this definition, where the function $\eta(\mathrm{PA}_{\mathcal{G}}(W_i), N(W_i))$ is constant with respect to fluctuation in the variables $N(W_i)$. Thus, taking $Z_i \coloneqq \mathrm{PA}_{\mathcal{G}'}(W_i)$ as the conditioning set satisfies Assumption 1, and the rest of our results hold, where the corresponding δ-perturbations in Definition 4 are given by

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i)) \right) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i | \, \mathrm{PA}_{\mathcal{G}}(V_i))$$

with shift function $s_i(\mathrm{PA}_{\mathcal{G}'}(W_i); \delta_i)$ that are parametric functions of causal parents in the modified graph $\mathcal{G}'$.

B.2.3. Domain-preserving parameterizations of shift

For both of the examples considered above, we did not need to restrict the magnitude of the additive change to $\eta(Z)$. However, in some cases, such as changing the variance of a conditional Gaussian, we have the restriction that $\eta_\delta(Z) = \eta(Z) + s(Z; \delta)$ must lie in the proper domain, e.g., we cannot consider a shift which causes the conditional variance to become negative. For a conditional Gaussian, we can consider unrestricted shifts in $\eta(Z)_1$, which controls the mean, because the mean has unrestricted domain. On the other hand, $\eta(Z)_2 = (-2\sigma^2(Z))^{-1}$ controls the variance, and must remain negative, such that $\eta(Z)_2 + s(Z; \delta)_2 < 0$ for the shifts we consider.

This can be resolved in one of two ways. First, one can consider parameterizations of $s(Z; \delta)$ which are guaranteed to preserve the correct domain with an additional constraint on the values of δ, such as the multiplicative shift below, which is sign-preserving for

$\delta > -1$

$$\eta_\delta(Z)_2 = \eta(Z)_2 + \underbrace{\delta\eta(Z)_2}_{s(Z;\delta)} = (1+\delta)\eta(Z)_2.$$

To handle the general case, at the expense of some additional complexity in the gradients of $s(Z;\delta)$, one can define the shifts as follows for parameters $\eta(Z)$ that have a lower bound L, with an equivalent formulation for shifts where the parameters have an upper bound, for any desired shift function $s'(Z;\delta)$

$$\eta(Z) + \underbrace{s'(Z;\delta) \cdot \text{sigmoid}(\gamma \cdot [(\eta(Z) + s'(Z;\delta)) - (L+\varepsilon)])}_{s(Z;\delta)}$$

where $\text{sigmoid}(\gamma\cdot(x-(L+\varepsilon)))$ is a smooth relaxation of the indicator function $\mathbf{1}\{x > L+\varepsilon\}$, for a sufficiently large temperature parameter $\gamma > 0$ and a small $\varepsilon > 0$. This transformation preserves the twice-differentiable nature of $s(Z;\delta)$. In practice, however, we typically evaluate the gradient of $s(Z;\delta)$ at $\delta = 0$, where $\eta(Z)$ does not lie at the boundary of allowable parameter space, such that we can consider simpler parameterizations like

$$\eta(Z) + \underbrace{s'(Z;\delta) \cdot \mathbf{1}\left\{\eta(Z) + s'(Z;\delta) > L + \varepsilon\right\}}_{s(Z;\delta)}$$

as long as ε is taken sufficient small such that $\eta(Z) > L + \varepsilon$ almost everywhere in $\mathbb{P}$.

B.3. Considerations and additional results for evaluation of the worst-case loss

In this section, we present additional results on the Taylor approximation and compare how the Taylor approximation compares to the reweighting approach in evaluation and worst-case optimization of the shifted loss.

- In Appendix B.3.1 we give a full treatment of how shift gradients and Hessians are estimated from samples, following Theorem 1.
- In Appendix B.3.2, we demonstrate in some cases, one does not need to estimate all of $\eta(Z)$, but only the parts of $\eta(Z)$ that is shifting.
- In Appendix B.3.3, we demonstrate that the second-order Taylor expansion is exact in a linear-Gaussian setting, which gives a conceptual connection between this work and that of Anchor Regression [Rothenhäusler et al., 2021], which considered a restricted type of additive shift intervention in a globally linear structural causal model.
- In Appendix B.3.4, we work out the expression for the shift gradient and Hessian when we condition on binary variables.

- In Appendices B.3.5 to B.3.7, we provide experiments that compare the variance of the importance sampling estimate $\hat{E}_{\delta,\text{IS}}$ (see (6)) to the variance of the Taylor estimate $\hat{E}_{\delta,\text{Taylor}}$ (see (7)) of the loss in a shifted distribution.

B.3.1. Algorithm for Estimation of Shift Gradients and Hessians

Here, we recall the form of the shift gradients and Hessians in Theorem 1, and demonstrate how to compute them in practice using a set of auxiliary regression functions fit to the validation data.

Theorem 1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 4, with intervened variables $\mathbf{W} = \{W_1, \ldots, W_m\}$ and shift functions $s_i(Z_i; \delta_i)$, where $\delta = (\delta_1, \ldots, \delta_m)$. Then the shift gradient is given by $\text{SG}^1 = (\text{SG}^1_1, \ldots, \text{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\text{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, T_i(W_i) \middle| Z_i\right)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\{\text{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, \varepsilon_{T_i|Z_i}\varepsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \varepsilon_{T|Z}\right] & i = j \\ \operatorname{cov}(\ell,\; D_{i,1}^\top \varepsilon_{T_i|Z_i}\varepsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i; \delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\varepsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Notation and Dimensions: Let $\mathbf{W} = \{W_1, \ldots, W_m\}$ denote the set of m intervened variables, and let $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$ denote the conditioning sets. Note that for a single $W_i \in \mathbb{R}^{d_{W_i}}$, we will generally have it that $Z_i \in \mathbb{R}^{d_Z}$, where d_W is the dimension of W (typically 1) and d_Z is the number of conditioning variables, and when considering n samples, W_i will be a matrix in $\mathbb{R}^{n \times d_W}$, and Z_i will be a matrix $\mathbb{R}^{n \times d_Z}$. The sufficient statistic $T_i(W_i)$ maps from $\mathbb{R}^{d_W}$ to $\mathbb{R}^{d_T}$, where d_T is the dimension of the sufficient statistic. For many common distributions, $T_i(W_i) = W_i$, the identity function. For others, like the conditional multi-variate Gaussian, $T_i(W_i) = [W_i, W_i W_i^\top]$, where $W \in \mathbb{R}^{d_W}$ and $W_i W_i^\top \in \mathbb{R}^{d_W \times d_W}$. In these cases, we squeeze $T_i(W_i)$ to be a single vector, so in this case $d_T = d_W + d_W^2$.

Auxiliary models: To estimate the shift gradients and Hessians, we first learn auxiliary predictive models, which are required for computing the relevant conditional covariances. For simplicity, we do not consider sample-splitting in the algorithm given below, but one could employ sample-splitting to learn these predictive models on an independent validation sample.

- For each W_i, we learn $\hat{\mu}_{W_i}(Z_i)$ as a regression model for $\mathbb{E}[T_i(W_i)|Z_i]$. Because $T_i(W_i)$ may have multiple dimensions, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}^{d_T}$.

- For each conditioning set Z_i, we learn $\hat{\mu}_\ell(Z_i)$ as a regression model for $\mathbb{E}[\ell|Z_i]$. Because the loss is one-dimensional, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}$.

We then construct the following, which are defined for each data point in the sample.

- For each W_i, we construct $\hat{\varepsilon}_{T_i|Z_i} \coloneqq T_i(W_i) - \hat{\mu}_{W_i}(Z_i)$, which is a vector of length d_{T_i}.
- For each conditioning set Z_i, for the loss ℓ, we construct $\hat{\varepsilon}_{\ell|Z_i} \coloneqq \ell - \hat{\mu}_\ell(Z_i)$, which is a real number.
- For each conditioning set Z_i, we compute $D_{i,1}(Z_i)$ as $\nabla_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a matrix of size $d_T \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.
- For each conditioning set Z_i, we compute $D_{i,2}(Z_i)$ as $\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a tensor of size $d_T \times d_{\delta_i} \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.

Estimating shift gradients The shift gradient and Hessian in Theorem 1 are expressed as conditional covariance. Since $\mathbb{E}[\mathrm{cov}(A, B|C)] = \mathbb{E}[\varepsilon_{A|C}\varepsilon_{B|C}]$ where $\varepsilon_{A|C} \coloneqq A - \mathbb{E}[A|C]$ and $\varepsilon_{B|C} \coloneqq B - \mathbb{E}[B|C]$, we can use the estimated conditional means above, to compute the shift gradient and Hessian. Suppose that we observe N samples, $n \in \{1, \ldots, N\}$. For each index $i \in [m] \coloneqq \{1, \ldots, m\}$,

$$\hat{\mathrm{SG}}^1_i = \frac{1}{N}\sum_{n=1}^{N} \hat{\varepsilon}^{(n)}_{\ell|Z_i} \cdot D_{i,1}(Z_i^{(n)})^\top \hat{\varepsilon}^{(n)}_{T_i|Z_i}$$

which yields a vector of length d_{δ_i}, and these are concatenated together for each i to yield the entire shift gradient. The shift Hessian is constructed block-wise, for each index $i, j \in [m] \times [m]$ as follows: If $i = j$, then we construct the corresponding $d_{\delta_i} \times d_{\delta_i}$ block as

$$\hat{\mathrm{SG}}^2_{i,i} = \frac{1}{N}\sum_{n=1}^{N} \hat{\varepsilon}^{(n)}_{\ell|Z_i} \cdot \left[\left(D_{i,1}(Z_i^{(n)})^\top \hat{\varepsilon}^{(n)}_{T_i|Z_i}\right)^{\otimes 2} - D_{i,2}(Z_i^{(n)})^\top \hat{\varepsilon}_{T_i|Z_i}\right]$$

where $v^{\otimes 2}$ denotes the outer product so that $v^{\otimes 2} = vv^\top$, and the transpose of $D_{i,2}$ refers to a transpose which has dimension $d_{\delta_i} \times d_{\delta_i} \times d_T$. On the other hand, if $i \neq j$ we have

$$\hat{\mathrm{SG}}^2_{i,j} = \frac{1}{N}\sum_{n=1}^{N} (\ell^{(n)} - \bar{\ell}) \cdot \left(D_{i,1}(Z_i^{(n)})^\top \hat{\varepsilon}^{(n)}_{T_i|Z_i}\right)\left(D_{j,1}(Z_j^{(n)})^\top \hat{\varepsilon}^{(n)}_{T_j|Z_j}\right)^\top$$

where $\bar{\ell}$ is the average value of ℓ in the validation sample.

B.3.2. Shifts where estimating all of $\eta(Z)$ is not necessary for estimating shift gradient and Hessian

The following example shows that when a shift occurs in an exponential conditional distribution with parameter $\eta(Z)$, we do not necessarily need to model all of $\eta(Z)$ in

order to compute the shift gradient and Hessian. In particular, we only need to model the parts of $\eta(Z)$ that shift. This is different from estimating the shifted loss using importance sampling, where $\eta(Z)$ needs to be evaluated to evaluate (5).

Example B.4. *Consider the distribution of W conditioned on variables Z that is a multi-variate Gaussian variable,*

$$W|Z = \mathcal{N}(\mu(Z), \Sigma(Z)),$$

for unknown functions μ, Σ. The sufficient statistic for the multivariate Gaussian distribution is $T(W) = (W, WW^\top)$ and the canonical parameter is $\eta(Z) = (\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1})$.[1] *The first component of $\eta(Z)$ is a signal-to-variance ratio and the second is the inverse covariance matrix. For a shift $(\delta, 0)$ that only affects the first component, we show that we do not need to model $\Sigma(Z)$, but only $\mu(Z)$. This is beneficial, since estimating a conditional covariance from data can be challenging, especially if W is high-dimensional.*

For $\delta \in \mathbb{R}^{d_W}$, let $s(Z;\delta) = (\delta, 0)^\top$, and suppose that we wish to estimate $\mathbb{E}_\delta[\ell]$ using (7). The derivative of s is given by

$$D_1 = \nabla_\delta^2 s(Z;\delta) = \left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right),$$

where the first block is a $d_W \times d_W$ diagonal matrix, and the second is a $d_W \times d_W^2$ matrix of zeros. The second derivative of s is $D_2 = 0$. Hence, using Theorem 1, the shift gradient is

$$\mathrm{SG}^1 = \mathbb{E}[D_1 \operatorname{cov}(\ell, (W, WW^\top)|Z)] = \mathbb{E}[\operatorname{cov}(\ell, W|Z)],$$

and

$$\begin{aligned} \mathrm{SG}^2 &= \mathbb{E}\left[D_1 \operatorname{cov}\left(\ell, \left(W - \mathbb{E}[W|Z], WW^\top - \mathbb{E}[WW^\top|Z]^{\otimes 2}\right)|Z\right) D_1^\top \right] \\ &= \mathbb{E}\left[\operatorname{cov}\left(\ell, \left(W - \mathbb{E}[W|Z]\right)^{\otimes 2}|Z\right) \right]. \end{aligned}$$

Conditional covariances can be computed by only residualizing one of the variables: $\mathbb{E}[\operatorname{cov}(A, B|C)] = \mathbb{E}[A(B - \mathbb{E}[B|C])]$. Thus, if we only residualize ℓ, we get

$$\mathrm{SG}^1 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z])W] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z]) \cdot (W - \mu(Z))^{\otimes 2}].$$

Therefore, given data from $\mathbb{P}$, we can estimate the shift gradients by plugging in estima-

[1] Or, more formally, $T(W) = \left(W, \operatorname{vec}(WW^\top)\right)$ and $\eta(Z) = \left(\sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\operatorname{vec}(\mu(Z))\right)$, where vec denotes the vectorization operation. For a detailed walk through of the exponential family parameterization of multivariate Gaussian distributions, see https://maurocamaraescudero.netlify.app/post/multivariate-normal-as-an-exponential-family-distribution/.

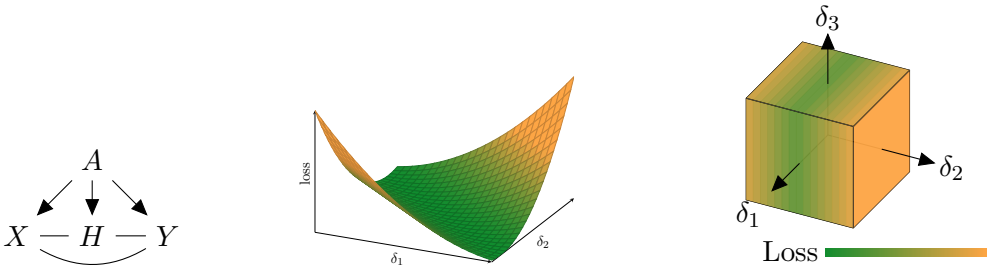


Figure B.2.: (Left) Graphical model assumed by (B.2). The undirected edges represent either any directed configuration of directed edges or the dependence structures arising due to an acyclic SCM [Bongers et al., 2021]. (Middle) Plotting $\mathbb{E}_\delta[(Y - \gamma^\top X)^2]$ as a function of $\delta \in \mathbb{R}^2$ for a fixed predictor γ. (Right) Plotting $\mathbb{E}_\delta[(Y - \gamma^\top X)^2]$ as a function of $\delta \in \mathbb{R}^3$, with the loss indicated by the color. The loss only varies with changes in δ_2 (corresponding in Lemma B.1 to $v_\gamma \propto (0, 1, 0)^\top$).

tors $\hat{\mu}(Z)$ *of* $\mathbb{E}[W|Z]$ *and* $\hat{L}(Z)$ *of* $\mathbb{E}[\ell|Z]$*. It follows that we do not need to model* $\Sigma(Z)$ *in order to estimate the shift gradients and Hessian at* $\delta = 0$*.*

The story is different for a reweighting based estimator that seeks to estimate $\mathbb{E}_\delta[\ell]$ *using importance sampling (see Section 3.1), where the weights are given by*

$$w_{\eta,\delta}(Z) = (W - \mu(Z))^\top \delta - \tfrac{1}{2}\delta^\top \Sigma(Z)\delta,$$

and hence estimating $w_{\eta,\delta}(Z)$ *requires estimation of* $\Sigma(Z)$*.*

B.3.3. The quadratic approximation is exact, for mean shifts in linear models

We now consider data generated by a linear model, and show that the shifted loss is a quadratic function of δ, meaning that the Taylor approximation $E_{\delta,\text{Taylor}}$ is globally exact. Suppose that data is sampled from a linear structural causal model, and a shift in mean occurs in an variable A that does not have any causal parents. In particular, let A have a normal distribution with mean μ and finite variance and let

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \varepsilon. \tag{B.2}$$

This is the model assumed by Rothenhäusler et al. [2021], and the corresponding graphical model is shown in Fig. B.2 (left). We consider the linear predictor $f_\gamma(X) = \gamma^\top X$ and the mean squared loss $\ell(f_\gamma(X), Y) = (Y - f(X))^2$. Due to the linearity of the model, the loss under a mean shift in A is quadratic [Rothenhäusler et al., 2021].

Lemma B.1. *Suppose* $A \sim \mathcal{N}(\mu, \Sigma)$ *and that* (X, Y, H) *are generated according to* (B.2)*. For* $\gamma \in \mathbb{R}^{d_X}$ *define* $\ell := (Y - \gamma^\top X)^2$*. Then there exist* $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ *such that for all*

shifts $\delta \in \mathbb{R}^{d_A}$*:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top \delta,$$

where $\mathbb{E}_\delta$ *corresponds to taking the mean in the distribution where* $A \sim \mathcal{N}(\mu + \delta, \Sigma)$*. Further* $u_{\mu,\gamma} = 0$ *if* $\mu = 0$*.*

Proposition B.2 elicits two properties of this linear model: First the loss is described by a quadratic function globally, i.e. also for very large δ. In Fig. B.2 (middle), we plot $\mathbb{E}_\delta[\ell]$ as a function of δ. We observe a 'valley' in the loss, in which the expected loss does not at all change with δ. This is a consequence of Lemma B.1, and particularly that if δ is orthogonal to both $u_{\mu,\gamma}$ and v_γ then $\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell]$. In higher dimensions $d_A > 2$, since $v_\gamma v_\gamma^\top$ has rank 1, the 'valley' persists in that the loss does not grow at all in $d_A - 2$ dimensions (or $d_A - 1$ if A has mean $\mu = 0$), see Fig. B.2 (right).

We now show that coefficients in the quadratic form in Lemma B.1 is equal to the shift gradient and Hessian. We use that the Gaussian distribution with known variance Σ can be parameterized as an exponential family with sufficient statistic $T(A) = \Sigma^{-1}A$ and parameter $\eta = \mu$.[2]

Proposition B.2. *Suppose* $A \sim \mathcal{N}(\mu, \Sigma)$ *and that* (X, Y, H) *are generated according to* (B.2)*. Then the shift gradient and Hessian are given by*

$$\mathrm{SG}^1 = \mathrm{cov}(\ell, \Sigma^{-1}A) \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathrm{cov}(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top})$$

and the loss under a mean shift of δ *in* A *is given by*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta,$$

where $\ell := (Y - \gamma^\top X)^2$ *and* $\mathbb{E}_\delta$ *corresponds to taking the mean in the distribution where* $A \sim \mathcal{N}(\mu + \delta, \Sigma)$*.*

This elicits a connection to anchor regression [Rothenhäusler et al., 2021]: Under the generative model (B.2) and using the quadratic loss $\ell = (Y - \gamma^\top X)^2$ for $\gamma \in \mathbb{R}^{d_X}$, they show that for any $\lambda \geq 0$, the worst-case loss $\mathbb{E}_\delta[\ell]$ over a set $\Delta = \{\delta | \delta\delta^\top \preceq \lambda \mathbb{E}[AA^\top]\}$ equals the objective $\ell_{\mathrm{AR}} = \mathbb{E}[\ell] + \lambda \mathbb{E}[\mathbb{E}[Y - \gamma^\top X | A]^2]$, which is computable from the observed distribution.

Because of Proposition B.2, ℓ_{AR} *also* equals the solution of the optimization problem (9) over the constraint set Δ. Therefore minimizing the anchor regression objective over γ or minimizing (9) over γ will lead to the same estimator. Since our proposed Taylor approximation in (9) does not assume linearity, one could use the approximation to extend the rationale of anchor regression of minimizing the worst-case loss to non-linear models. This however comes at the cost of not optimizing the exact worst-case loss, but rather an approximation, whose quality is given by Theorem 2. Further, this would involving a minimax problem, minimizing (9) over models f, and there are questions, such as convexity and tractability, which would need to be solved.

[2] It can also be parameterized as $T(A) = \Sigma^{-1/2}A, \eta = \Sigma^{-1/2}\mu$, which would yield the same result.

B.3.4. Estimating the shift gradient and Hessian for conditional on binary variables

To build intuition for the shift gradient and Hessian, we here give an example where we condition on variables Z that take a finite number of values and write out explicit expressions for the shift gradient and Hessian. However, we emphasize, that in most practical scenarios, one will not have to work out the shift gradient and Hessian explicitly, but can simply estimate them as covariances from the data (Theorem 1).

Example B.5 (Shift Function of Discrete Parents)**.** *Consider a conditional distribution $W|Z$ where Z takes values in a finite set $\mathcal{Z}$. This is for instance the case if $Z = (Z_1, \dots, Z_d)$ where each Z_i is binary, so $|\mathcal{Z}| = 2^d$. Instead of a shift $\eta(Z) + \delta$, where the parameter increases by the same amount for all values of Z, we may consider a shift $\eta(Z) + s(Z;\delta)$ where $s(Z;\delta) = \sum_{z\in\mathcal{Z}} \delta_z 1_{Z=z}$, meaning that the shift is different in each category Z. Since $\eta(Z)$ only takes a finite number of variables, this shift corresponds to an arbitrary change in $\eta(Z)$.*

$s(Z;\delta)$ is a differentiable function in δ, and if $d_T = 1$ the shift gradient is a (1×2^d)-row vector, $\nabla_\delta s(Z;\delta) = (1_{Z=z})_{z\in\mathcal{Z}}$, and the shift Hessian vanishes, $\nabla^2_\delta s(Z;\delta) = 0$. Enumerating $\mathcal{Z} = \{1, \dots, 2^d\}$, the i'th entry in the shift gradient becomes

$$(\mathrm{SG}^1)_i = \mathbb{E}\left[1_{Z=i} \operatorname{cov}\left(\ell, T(W)\middle| Z\right)\right] = \mathbb{P}(Z=i)\operatorname{cov}(\ell, T(W)|Z=i),$$

and the i,j'th entry of the shift Hessian becomes 0 if $j \neq i$ and else

$$(\mathrm{SG}^2)_{i,i} = \mathbb{E}\left[1_{Z=i} \operatorname{cov}_\delta\left(\ell, \varepsilon_{T|Z}^{\otimes 2}\middle| Z\right)\right] = \mathbb{P}(Z=i)\operatorname{cov}(\ell, \varepsilon_{T|Z}^{\otimes 2}|Z=i).$$

Consider for example the case where both W and Z are binary. Then $T(W) = W$ and $s(Z;\delta) = 1_{Z=0}\delta_0 + 1_{Z=1}\delta_1$ and $s^{(1)} = (1_{Z=0}, 1_{Z=1})$ and $s^{(2)} = 0$. The conditional covariance can be evaluated by residualizing only one of the variables, $\mathbb{E}[\operatorname{cov}(A, B|C)] = \mathbb{E}[A(B - \mathbb{E}[B|C])]$, so we can chose to residualize only W (for SG^1) or $(W - \mathbb{E}[W|Z=i])^2$ (for SG^2). Finally, if we let $p_i = \mathbb{P}(W=1|Z=i)$ and use that $\mathbb{E}[W|Z=i] = p_i$ and $\mathbb{E}[(W-p_i)^2|Z=i] = \operatorname{var}(W|Z=i) = p_i(1-p_i)$, we get that

$$\mathrm{SG}^1 = \mathbb{E}\left[\begin{pmatrix} p_0 \cdot \ell \cdot (W - p_0) \\ p_1 \cdot \ell \cdot (W - p_1) \end{pmatrix}\right],$$

and

$$\mathrm{SG}^2 = \mathbb{E}\left[\begin{pmatrix} \ell p_0\{(W-p_0)^2 - p_0(1-p_0)\} & 0 \\ 0 & \ell p_1\{(W-p_1)^2 - p_1(1-p_1)\} \end{pmatrix}\right].$$

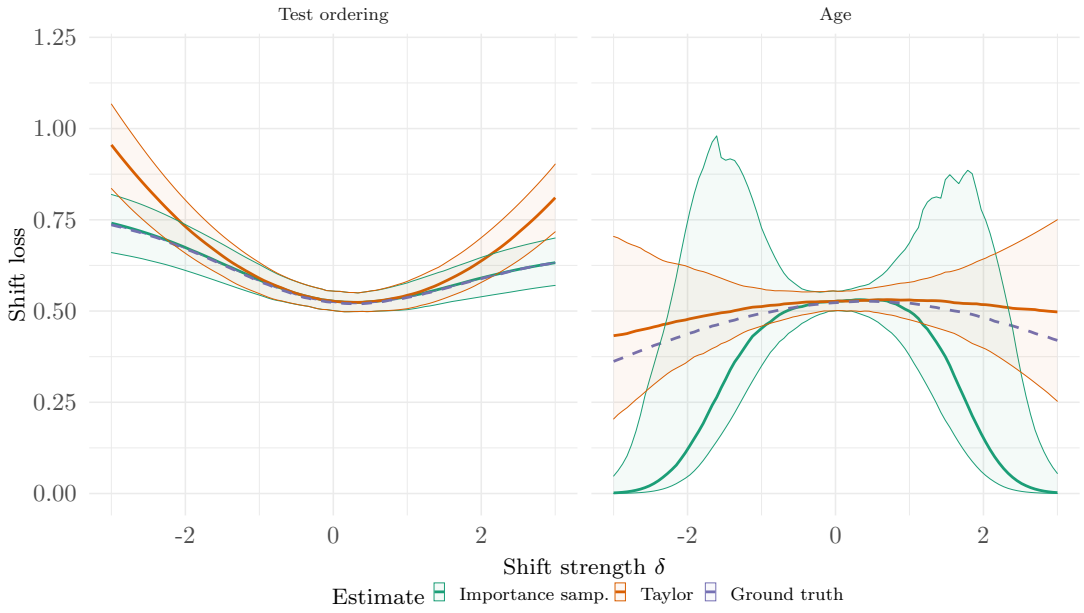


Figure B.3.: We plot the mean and confidence intervals of $\hat{E}_{\delta,\text{Taylor}}$ and $\hat{E}_{\delta,\text{IS}}$ when the shifted loss as in the lab test ordering example Example 1. (Left) We consider a shift in the logits of ordering lab tests from $\eta(Z)$ to $\eta(Z) + \delta_0$. (Right) We consider a shift in the mean of Age. In the observed distribution $\eta = \mu/\sigma = 0$ and we shift to a mean of $\eta = \delta$.

B.3.5. Comparison of variance of reweighting and Taylor estimates in the lab ordering example

To compare the bias and variance of the Taylor and the importance sampling estimates of the shifted loss, we simulate data from the following, artificial, generative model (which is the same generative model that was used to construct the loss landscape in Fig. 1 (right)).

$$\begin{aligned}
\text{Age} &\sim \mathcal{N}(0, 0.5^2) \\
\mathbb{P}(\text{Disease} = 1 | \text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\
\mathbb{P}(\text{Order} = 1 | \text{Disease}, \text{ Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\
\text{Test Result} | \text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1)
\end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero.

We consider either a shift in the logits of ordering lab tests $\eta_\delta(Z) = \eta(Z) + \delta$ (Fig. B.3 left) or a mean shift in the Gaussian distribution of age $\eta_\delta = \delta$ (Fig. B.3 right). For each δ in a grid, we compute estimates $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ of the loss under a shift of size δ, We repeat this $n = 1{,}000$ times, and plot the mean and point-wise prediction intervals (the pointwise 0.05 and 0.95 quantiles) for $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$. We also simulate ground

truth data from $\mathbb{P}_\delta$, to compute the actual loss under shift.

For shifts in the binary variable (Fig. B.3, left), both estimates capture the loss well for small shifts, but as δ gets larger, the quadratic approximation increasingly deviates from the true mean; the importance sampling estimate remains very close to the ground truth shifted loss. On the contrary, for the Gaussian mean shift (Fig. B.3, right), the importance sampling weights are ill-behaved, and the variance dramatically increases as δ becomes larger. This supports the intuition, that while importance sampling tends to work well for binary variables, the variance can be large in continuous distributions, such as the Gaussian distribution.

B.3.6. Comparison of theoretical variance of reweighting and Taylor estimates

Example B.6. *To demonstrate the reduction in variance obtained from using the Taylor approximation of the importance weights, we consider a simple example where* $\mathbb{P}(X) \sim \mathcal{N}(0,1)$ *and* $\mathbb{P}_\delta(X) \sim \mathcal{N}(\delta,1)$ *and we wish to estimate* $\mathbb{E}_\delta[\ell(X)]$ *for some loss function* $\ell(X)$.[3] *The importance sampling weights are given by* $w_\delta(X) = \exp(-\frac{1}{2}\delta^2 + X\cdot\delta)$, *and the shift gradient and Hessians are* $\mathrm{SG}^1 = \mathbb{E}[\ell(X)X]$ *and* $\mathrm{SG}^2 = \mathbb{E}[\ell(X)X^2]$.

Therefore samples $X_1,\dots,X_n$ *from* $\mathbb{P}$ *consider the estimators, for any loss function* $\ell(X)$, *two estimators of* $\mathbb{E}_\delta[\ell]$ *are*

$$\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^n w_\delta(X_i)\ell(X_i) \quad \textit{and} \quad \hat{\mu}_{Taylor} = \frac{1}{n}\sum_{i=1}^n \ell(X_i) + \delta\cdot\ell(X_i)X_i + \tfrac{1}{2}\delta^2\ell(X_i)X_i^2,$$

and the variances of the estimators are

$$\mathrm{var}(\hat{\mu}_{IS}) = \frac{\mathbb{E}[\{\ell(X+2\delta)\}^2]}{n}\exp(\delta^2)$$
$$\mathrm{var}(\hat{\mu}_{Taylor}) = \frac{\mathrm{var}\left(\ell(X) + \delta X\ell(X) + \frac{1}{2}\delta^2X^2\ell(X)\right)}{n}.$$

The variance of $\hat{\mu}_{Taylor}$ *grows like* δ^4 *and the variance of* $\hat{\mu}_{IS}$ *grows exponentially fast (unless* $\mathbb{E}[\{\ell(X+2\delta)\}^2]$ *also diminishes exponentially fast, which is generally not the case), and so except for small* δ, *the variance of the importance sampling estimator will be orders of magnitude larger than the variance of the estimator using the Taylor approximation. While,* $\hat{\mu}_{IS}$ *is an unbiased estimator of* $\mathbb{E}_\delta[\ell(X)]$ *and* $\hat{\mu}_{Taylor}$ *is a biased, the overall mean squared error will be smaller for the Taylor approximation, unless the bias of the Taylor approximation also grows exponentially.*

For the sake of analysis, consider the simple example $\ell(X) = X$. *In this case, the Taylor estimate is unbiased because* $\mathbb{E}_\delta[X] = \delta$ *is a linear function of* δ, *so the quadratic*

[3] In practice one would not use importance sampling estimation for such a simple shift, but use other approaches, such as analytically work out an estimate of $\mathbb{E}_\delta[\ell]$.

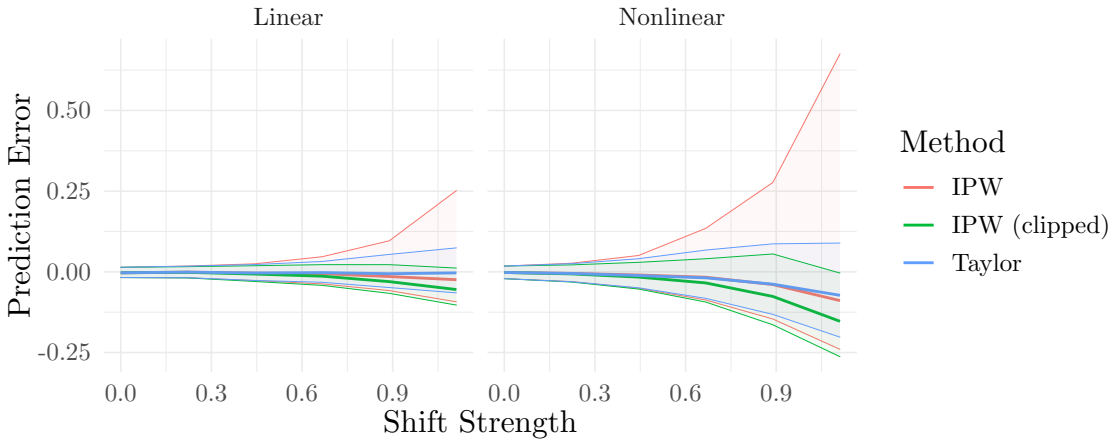


Figure B.4.: Median and quantiles of the error in predicting $\mathbb{E}_\delta[\ell]$ under a shift δ.

approximation is adequate. Further, the variances are given by

$$\mathrm{var}(\hat{\mu}_{IS}) = \frac{\exp(\delta^2)(1+4\delta^2) - \delta^2}{n} \quad \textit{and} \quad \mathrm{var}(\hat{\mu}_{Taylor}) = \frac{1 + 5\delta^2 + \frac{15}{4}\delta^4}{n}.$$

In particular, the variance of the importance sampling estimate grows like $\exp(\delta^2)$ *while that of the Taylor estimate grows like* δ^4.

B.3.7. Comparison of variance of reweighting and Taylor estimates in a simple synthetic example

In this experiment, we compare the variance of importance sampling and Taylor estimates in a simple synthetic example. We simulate data from $\mathbb{P}$ where $X \in \mathbb{R}^3$ and $Y \in \mathbb{R}^1$ depend either linearly or quadratically on $W \in \mathbb{R}^3$,

$$W \sim \mathcal{N}(0, \mathrm{Id}_3) \qquad \text{and} \qquad \begin{pmatrix} X \\ Y \end{pmatrix} = (\mathrm{Id}_4 - B)^{-1} M (W + \alpha(W \odot W) + \varepsilon),$$

where $\odot$ refers to entrywise multiplication, $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_4)$, α is either 0 (linear) or $\frac{1}{2}$ (nonlinear) and

$$B := \begin{pmatrix} 2 & 1 & 0 & 1 \\ 2 & 2 & 0 & 3 \\ 3 & 3 & 0 & 2 \\ 4 & 2 & 4 & 0 \end{pmatrix} \qquad \text{and} \qquad M := \begin{pmatrix} 2 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 2 & 0 \\ 4 & 1 & 1 \end{pmatrix}.$$

On the simulated data from $\mathbb{P}$, we then fit a linear predictor $f(X)$ of Y, and consider a shift in the mean of W from $\mathbb{P}(W) \sim \mathcal{N}(0, \mathrm{Id}_3)$ to $\mathbb{P}_\delta(W) \sim \mathcal{N}(\delta, \mathrm{Id}_3)$, where $\delta = [s, s, s]^\top$ for some shift strength $s > 0$. We then compute the shift gradient $\mathrm{SG}^1 = \mathrm{cov}(\ell, W)$ and Hessian $\mathrm{SG}^2 = \mathrm{cov}(\ell, WW^\top)$, and approximate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\mathrm{Taylor}}$ (see (7)). In the linear

data, the Taylor approximation is exact (see Appendix B.3.3), such that any prediction error can be attributed to finite-sample fluctuation, whereas both model misspecification and finite-sample fluctuation contribute to the error in the nonlinear setting.

Similarly, we estimate $\mathbb{E}_\delta[\ell]$ by importance sampling, $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(W)\ell] \approx \frac{1}{n}\sum w_\delta(W)\ell$, where $w_\delta(W) = \frac{\mathbb{P}_\delta(W)}{\mathbb{P}(W)} = \delta^\top W - \frac{1}{2}\delta^\top\delta$, and compare this to ground truth data sampled from $\mathbb{P}_\delta$; we do the same for an importance sampling estimator with weights 'clipped' at the 99% quantile.

We compare the predicted loss $\mathbb{E}_\delta[\ell]$ by actually simulating data from $\mathbb{P}_\delta$ and evaluating $\mathbb{E}_\delta[\ell]$ (where ℓ is still the model trained on data from $\mathbb{P}$). We then compute the prediction error, as the difference $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\text{Taylor}}$ or $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\text{IS}}$.

For a number of different shift strengths s, we repeat this procedure $M = 1{,}000$ times, and in Fig. B.4 we plot the median and a confidence interval defined by the 2.5 and the 97.5% quantiles of the prediction error.

In the linear case, bot the importance sampling and the Taylor approximation retains a median error close to 0, with the variance of $\hat{E}_{\delta,\text{IS}}$ being larger than $\hat{E}_{\delta,\text{Taylor}}$. The clipped importance sampling estimate has a smaller variance than that of ordinary importance sampling, though the median deviates further from 0, and the variance is not smaller than that of the Taylor estimate.

In the non-linear cases, all three models underestimate the shifted loss. For $\hat{E}_{\delta,\text{Taylor}}$, this happens because as the mean of W shift, the mean shift is amplified by the non-linearity, such that the quadratic approximation of the loss is an underestimate. While the variance of the clipped importance sampling is smaller than the variance of the ordinary importance sampling estimate and comparable to the variance of the Taylor estimate, this prediction is further from 0 than the Taylor estimate.

Since importance sampling methods are known to produce very large outliers, the use of the median and quantiles, as opposed to the mean an confidence intervals based on the standard deviation, is favouring importance sampling; the Taylor method looks even more favourable if we instead plot the mean and standard deviations.

B.4. Limitations of worst-case conditional subpopulation shift for defining plausible robustness sets

For the example in Section 4.1, we can contrast the type of shift we consider with the worst-case $(1-\alpha)$-conditional subpopulation shift considered by Subbaswamy et al. [2021].

In this section, we will make the following points: First, worst-case conditional $(1-\alpha)$-subpopulation shifts can be too pessimistic, with even moderate values of α leading to implausible conditional distributions. Second, we will argue that parametric robustness sets enable more fine-grained control over the set of plausible shifts, leading to more informative estimates of worst-case risk. Overall, we argue that the two approaches are complementary, with different strengths.

Before we proceed, we define a conditional $(1-\alpha)$ subpopulation shift. A $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ is defined by a weighting

function $h : \mathcal{O} \times \mathcal{Y} \mapsto [0, 1]$, which has the property that $\mathbb{E}[h(O, Y)|Y] = 1 - \alpha$ for all values of Y. This can be used to construct a worst-case objective, which measures the worst-case loss under such a shift:

$$\begin{aligned} \sup_{h:\{0,1\}^2 \mapsto [0,1]} \quad & \frac{1}{(1-\alpha)} \mathbb{E}[h(O, Y)\mu(O, Y)] \\ \text{s.t.} \quad & \mathbb{E}[h(O, Y)|Y = y] = 1 - \alpha, \quad \text{for } y \in \{0, 1\} \end{aligned} \tag{B.3}$$

where $\mu(O, Y) \coloneqq \mathbb{E}[\ell(Y, f)|O, Y]$, for a predictor f and loss ℓ. This has the effect of leaving the distribution $\mathbb{P}(Y)$ untouched, while changing the conditional distribution $\mathbb{P}(O|Y)$. Throughout this section, we will use the same predictor $f(O, L)$ described in Section 4.1. The rest of this section is structured as follows:

In Appendix B.4.1, we derive the feasible set of conditional distributions $\mathbb{P}(O|Y)$ implicitly considered by this objective in the simple generative model of Section 4.1, which only involves variables O, L and Y. We do so by showing that (for discrete O, Y), maximizing (B.3) over h is equivalent to solving a linear program, where we can characterize the constraints on h exactly, and translate them into constraints on $\mathbb{P}(O = 1|Y = 1), \mathbb{P}(O = 1|Y = 0)$. Here, we show that the resulting feasible set is quite large, even for moderately large subpopulations. In particular, whenever $(1-\alpha) < \min\{\mathbb{P}(O = 1|Y = 0), \mathbb{P}(O = 0|Y = 1)\}$, all conditional distributions are possible.

In Appendix B.4.2, we derive the value of h that maximizes (B.3), and show that, as we vary α, the worst-case shift is always in the same "direction" probability space: Healthy patients ($Y = 0$) are tested more, and sick patients ($Y = 1$) are tested less, and for $\alpha < 0.27$, the worst-case subpopulation shift is the (unrealistic) scenario where healthy patients are always tested, and sick patients are never tested.

In Appendix B.4.3, we illustrate how this type of behavior can be avoided with our approach. We first give a parameterized shift function $s(Z; \delta_0, \delta_1)$ such that we can reach any conditional distribution of $\mathbb{P}(O|Y)$, for sufficiently large values of δ_0, δ_1. We then demonstrate how an iterative process might play out with domain experts, where we consider different constraint sets until we find a constraint set that contains plausible shifts.

B.4.1. Feasible conditional subpopulations in Section 4.1

For the simple example in Section 4.1, we give a self-contained derivation of the feasible region for $1 - \alpha$ conditional subpopulations in the distribution $\mathbb{P}(O|Y)$. The advantage of working with this simple generative model is that the conditional distribution can be described by only two numbers, $\mathbb{P}(O = 1|Y = 1)$ and $\mathbb{P}(O = 1|Y = 0)$, and so we can visualize the resulting conditional distribution.

Because O, Y are discrete, the worst-case subpopulation in this simple example can be solved via a linear program, for a fixed α. We have an optimization problem in two variables, since $h_{11}\mathbb{P}(O = 1|Y = 1) + h_{01}\mathbb{P}(O = 0|Y = 1) = 1 - \alpha$, and likewise for h_{10}, h_{00}, where $h_{ij} = h(O = i, Y = j)$. We also have the constraint that each variable must live in $[0, 1]$. Meanwhile, the loss to maximize is a linear function, as an

expectation of $\mathbb{E}[h(O,Y)\mu(O,Y)]$, where $\mu(O,Y)$ takes on four possible values, where we write $p_{ij} = \mathbb{P}(O = i|Y = j)$, and μ_{ij} similarly.

$$\begin{aligned} \max_{h \in \mathbb{R}^{2\times 2}} \quad & h_{00}\mu_{00} + h_{10}\mu_{10} + h_{01}\mu_{01} + h_{11}\mu_{11} \\ \text{s.t.,} \quad & h_{11}p_{11} + h_{01}(1-p_{11}) = 1-\alpha \\ & h_{10}p_{10} + h_{00}(1-p_{10}) = 1-\alpha \\ & 0 \le h_{ij} \le 1, \forall i,j \end{aligned} \tag{B.4}$$

This linear program is simple enough to solve by hand, and we will do here to build intuition. In this section, we begin by characterizing the feasible region of h, and then translating that into a feasible region for $\mathbb{P}_h(O|Y)$, which we can plot in two dimensions.

Characterizing feasible values of h: Here, we focus on characterizing the feasible set that h can lie in, as a way of characterizing the feasible set for $\mathbb{P}(O|Y)$. From the constraints, we can write that

$$\begin{aligned} h_{11}p_{11} + h_{01}(1-p_{11}) = 1-\alpha \qquad &\implies \qquad h_{01} = \frac{1-\alpha-h_{11}p_{11}}{1-p_{11}} \\ h_{10}p_{10} + h_{00}(1-p_{10}) = 1-\alpha \qquad &\implies \qquad h_{00} = \frac{1-\alpha-h_{10}p_{10}}{1-p_{10}} \end{aligned}$$

There are only two constraints on h_{11}: Those directly imposed by $0 \le h_{11} \le 1$, and those which are imposed by the equality constraint with h_{01} and the fact that $0 \le h_{01} \le 1$. For the latter, with some algebra we can write that

$$0 \le \frac{1-\alpha-h_{11}p_{11}}{1-p_{11}} \le 1 \qquad \implies \qquad \frac{p_{11}-\alpha}{p_{11}} \le h_{11} \le \frac{1-\alpha}{p_{11}}$$

So that the constraints on h_{11} become

$$\max\left\{0, \frac{p_{11}-\alpha}{p_{11}}\right\} \le h_{11} \le \min\left\{1, \frac{1-\alpha}{p_{11}}\right\} \tag{B.5}$$

which recovers our intuition that if $\alpha = 0$, it must be that $h_{11} = 1$ and $h_{01} = 1$.

Bounding feasible values of $\mathbb{P}_h(O|Y)$ The parameters h can be understood as importance weights whose expectation is $1-\alpha$ instead of 1, that reweight $\mathbb{P}$ to a new distribution $\mathbb{P}_h$ when appropriately normalized. To compute conditional probabilities $\mathbb{P}_h(O = i|Y = j)$ under the new distribution, we can compute the expectation of $\mathbf{1}\{O = i, Y = j\}$, and normalize by $\mathbb{P}(Y = j)$.

$$\begin{aligned} & \mathbb{P}_h(O=i, Y=j) = \frac{1}{1-\alpha}\mathbb{E}[h(O,Y)\mathbf{1}\{O=i,Y=j\}] = \frac{h_{ij}}{1-\alpha}\mathbb{P}(O=i,Y=j) \\ \implies & \mathbb{P}_h(O=i|Y=j) = \frac{h_{ij}}{1-\alpha}\mathbb{P}(O=i|Y=j) \end{aligned}$$

where the implication follows from the fact that $\mathbb{P}_h(Y) = \mathbb{P}(Y)$. This allows us to

translate bounds on h_{ij} directly into bounds on $\mathbb{P}_h(O = i|Y = j)$. Making use of (B.5), we can write that

$$\max\left\{0, \frac{p_{11} - \alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1-\alpha} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{1, \frac{1-\alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1-\alpha}$$

which yields

$$\max\left\{0, \frac{p_{11} - \alpha}{1-\alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{\frac{p_{11}}{1-\alpha}, 1\right\}$$

We can apply a similar logic to h_{10}, which is identical except for p_{11} being replaced by p_{10}, yielding

$$\max\left\{0, \frac{p_{10} - \alpha}{1-\alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 0) \leq \min\left\{\frac{p_{10}}{1-\alpha}, 1\right\}$$

Visualizing the constraint set: Figure B.5 gives feasible conditional distributions under different values of α. We can observe that when $\alpha = 0.8$, all conditional distributions are feasible, including the distribution where $\mathbb{P}(O = 1|Y = 0) = 1$ and $\mathbb{P}(O = 1|Y = 1) = 0$, representing the case where every healthy patient gets tested, and no sick patients receive a test. This is generally possible in this example whenever $1 - \alpha < \min\{\mathbb{P}(O = 1|Y = 0), \mathbb{P}(O = 0|Y = 1)\}$, as it permits the following subpopulation function, which yields this result.

$$h(O = o, Y = y) = \frac{1-\alpha}{\mathbb{P}(O = o|Y = y)} \mathbf{1}\{o \neq y\}$$

B.4.2. Worst-case conditional subpopulation shifts

Given the constraint set which describes the feasible set of conditional distributions under the $(1 - \alpha)$-conditional subpopulation objective, we can derive the worst-case conditional distribution. Here, since Y, O are both binary, the expected loss under a new distribution $\mathbb{P}_h$ is given by

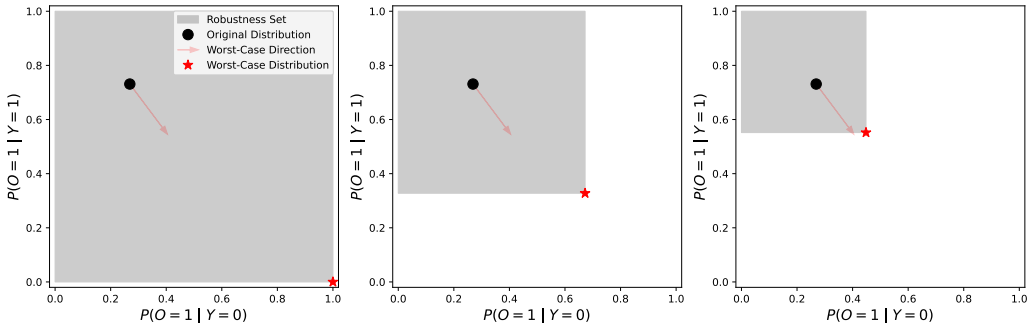
$$\mathbb{E}_h[\ell] = \sum_{y,o} \mu(o, y)\mathbb{P}_h(O = o|Y = y)\mathbb{P}(Y = y)$$

which we can write in terms of the constrained probabilities $\mathbb{P}_h$ as follows, where $q_{11} \coloneqq \mathbb{P}_h(O = 1|Y = 1)$ and $q_{10} \coloneqq \mathbb{P}_h(O = 1|Y = 0)$

$$\mathbb{P}(Y = 1)[\mu(1, 1)q_{11} + \mu(0, 1)(1 - q_{11})] + \mathbb{P}(Y = 0)[\mu(1, 0)q_{10} + \mu(0, 0)(1 - q_{10})]$$

which also gives us a direction in which the loss is maximized, since the loss is given by

$$\mathbb{E}_h[\ell] = q_{11} \cdot \mathbb{P}(Y = 1) \cdot (\mu(1, 1) - \mu(0, 1)) + q_{10}\mathbb{P}(Y = 0) \cdot (\mu(1, 0) - \mu(0, 0)) + C \quad \text{(B.6)}$$



(a) $(1-\alpha) = 0.2$ (b) $(1-\alpha) = 0.4$ (c) $(1-\alpha) = 0.6$

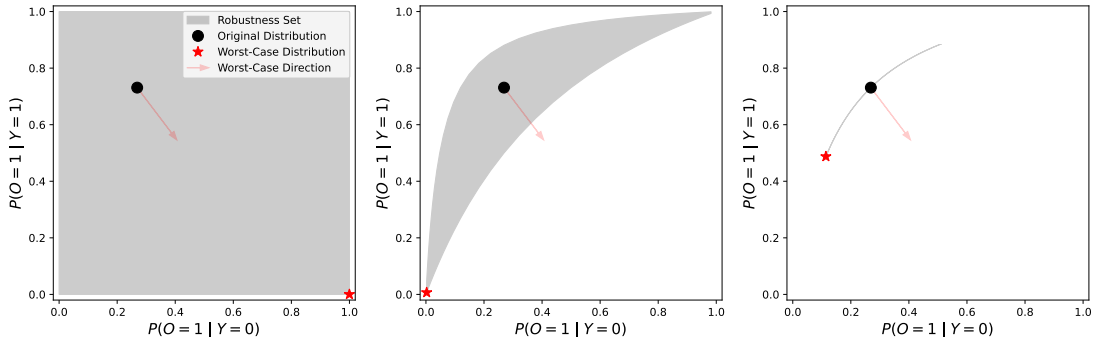
Figure B.5.: Feasible sets, worst-case directions, and worst-case solutions for a $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ for differing values of α. Worst-case directions are computed using (B.6), as unit-norm vectors re-scaled to fit in the plot, and the colored dots give the worst-case solutions, all of which lie in the lower-right corner of the constraint set. The original conditional distribution is given by the black dot.

where $C = \mathbb{P}(Y = 1)\mu(0,1) + \mathbb{P}(Y = 0)\mu(0,0)$. Since q_{11}, q_{10} can be optimized independently, the worst-case solution is given by taking the maximum value of q_{11} if $\mu(1,1) > \mu(0,1)$ and the minimum value if $\mu(1,1) < \mu(0,1)$, and likewise taking the maximum value of q_{10} if $\mu(1,0) > \mu(0,0)$, and the minimum value otherwise. If $\mu(1,1) = \mu(0,1)$ or $\mu(1,0) = \mu(0,0)$, then the objective is unaffected by the choice of q_{11} or q_{10} respectively.

Visualizing the worst-case conditional distributions The worst-case directions on the probability scale, and the resulting worst-case conditional distribution obtained by solving (B.4), are given in Fig. B.5. The red line arrow visualizes the direction from (B.6), and the worst-case distribution is the point which is furthest in this direction in the constraint set. Here, we are finding the worst-case accuracy of the same predictive model $f(O, L)$ described in Section 4.1. We can observe that the worst-case loss is obtained by seeking to reverse the correlation between Y and O, decreasing the probability that a sick patient $(Y = 1)$ gets a test ordered, and increasing the probability that a healthy patient $(Y = 0)$ gets a test ordered.

B.4.3. Iterating with domain experts to define realistic parametric robustness sets

In the previous sections, we saw that $(1-\alpha)$-conditional subpopulation shift does not always produce realistic worst-case conditional distributions. Moreover, given only the parameter α, there is limited ability to control the nature of the resulting worst-case conditional distribution $\mathbb{P}(O|Y)$. In this section, we contrast this limitation with the finer-grained control enabled by considering parametric robustness sets. In particular, we argue that parametric shifts allow for end-users to customize robustness sets, ruling out shifts that represent unrealistic changes.



(a) (b) (c)

Figure B.6.: Each figure shows the set of conditional probability distributions ("CPDs") $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case distribution (given by the red star) for the 0–1 loss. In this example, the expected loss under $\mathbb{P}_\delta$ is a linear function of the two conditional probabilities (see Appendix B.4.2), where the loss increases along the red arrow. (a) captures (nearly) all conditional probability distributions, with Δ_0, Δ_1 unconstrained. (b) shows a set of CPDs with Δ_0 unconstrained, and $\Delta_1 = [-1, 1]$, with resulting worst-case accuracy of 50%. (c) shows a more restrictive set of shifts, where $\Delta_0 = [-1.05, 1.05], \Delta_1 = \{0\}$. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.

In practice, we imagine that the following iterative process could be a useful tool in model development: (i) Define a class of shifts with an appropriate $s(Z; \delta)$ and constraint set Δ, and search for a worst-case shift δ. (ii) Present to domain experts **both** the worst-case shift δ (in terms of summary statistics of the resulting distribution $\mathbb{P}_\delta$) alongside the associated estimate of the worst-case loss. For instance, report both the worst-case loss, as well as corresponding rate of testing among sick and healthy patients. (iii) If the shift itself is unrealistic, further the constrain parameter set or shift function, and repeat the process.

In Fig. B.6, we give a concrete example. Each sub-figure shows the set of conditional probability distributions $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case conditional distribution (given by the red star) for the 0–1 loss. Recall that we use the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 controls a general increase or decrease in testing, while δ_1 controls a shift in the testing rate for only sick patients, and allows for a different change in the testing rate of sick vs healthy patients.

Iteration 1: We might imagine starting with a relatively unconstrained robustness set, where δ_0 and δ_1 are unconstrained. Figure B.6a shows the resulting robustness set of conditional distributions, and finds a shift with with a worst-case accuracy of 16%, compared to accuracy of 75% on the original distribution. However, the corresponding δ-perturbation $\mathbb{P}_\delta$ is unrealistic, where all healthy patients (and no sick patients) are tested. Luckily, because we have parameterized the shift, we can constrain the robustness set

to exclude these types of results.

Iteration 2: A benefit of our approach is that we can refine the robustness set, with this type of feedback in mind. In Fig. B.6b, we restrict the support of δ_1 to $[-1, 1]$, to avoid large changes in the relative probability of testing sick vs healthy patients. Here, the resulting worst-case accuracy is much higher (50%), but the corresponding worst-case conditional probability distribution is perhaps still unrealistic: No patients undergo laboratory testing at all!

Iteration 3: Finally, we consider only shifts that affect all patients in a similar way, generally raising or lowering the conditional probability of a lab test, represented by shifts in δ_0 alone. This may correspond to a more realistic scenario where (in a new hospital) laboratory testing use is more or less constrained. Additionally, we can specify that this shift should decrease testing rates by at most 20%, which translates directly into a lower-bound on δ_0.[4] Figure B.6c shows the resulting robustness set of distributions, where the worst-case shift may seem more plausible: A reduction in testing rates for both populations. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.

B.5. CelebA: Experiment details and additional results

In this section, we give details of the computer vision experiment in Section 4.2.

B.5.1. Details for the experiment

Creating the training distribution To construct the training distribution $\mathbb{P}$, we use the conditional GAN in Kocaoglu et al. [2018]. In particular, we use their CausalBEGAN, which is an extends the boundary equillibrium GAN [Berthelot et al., 2017] to also take attributes as inputs. We train the CausalBEGAN using the default hyper parameters in the implementation provided by Kocaoglu et al. [2018], available under the MIT license. The model is trained for 250,000 iterations on a single GPU, taking around approximately 16 hours.

Similar to Kocaoglu et al. [2018], we use the CelebA dataset [Liu et al., 2015], which contains approximately 200,000 images of faces, along 40 binary attributes. Of those, we use the following attributes 9 attributes {Male, Young, Wearing Lipstick, Bald, Mustache, Eyeglasses, Narrow Eyes, Smiling, Mouth Slightly Open}. The CelebA dataset is licensed for non-commercial research purposes only, and consists of publicly available images of celebrities, which were collected from the internet. Although the data set has been widely used, Liu et al. [2015] do not make any mention of consent by the individuals to have the images included in the data set, and it is therefore likely that those celebrities did not provide consent.

[4] In Proposition B.1, we prove that for binary random variables with a shift $\eta(Z) + \delta$, there is a one-to-one mapping between a new marginal distribution ($\mathbb{P}(O = 1)$ in this case) and the value of the parameter δ.

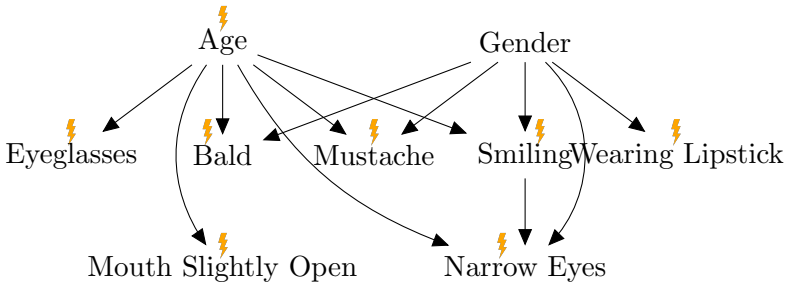


Figure B.7.: Causal graph over attributes, where lightning bolts indicate changes in mechanisms. Also displayed in Fig. 4.

Training distribution over attributes For the training distribution, we simulate binary attributes according to the structural causal model in Fig. 4 (for convenience also copied to Fig. B.7), where the model parameters are

$$
\begin{aligned}
\mathbb{P}(\text{Young} = 1) &= \sigma(0.0) \\
\mathbb{P}(\text{Male} = 1) &= \sigma(0.0) \\
\mathbb{P}(\text{Eyeglasses} = 1|\text{Young}) &= \sigma(0.0 - 0.4 \cdot \text{Young}) \\
\mathbb{P}(\text{Bald} = 1|\text{Young, Male}) &= \sigma(-3.0 + 3.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Mustache} = 1|\text{Young, Male}) &= \sigma(-2.5 + 2.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Smiling} = 1|\text{Young, Male}) &= \sigma(0.25 - 0.5 \cdot \text{Male} + 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Wearing Lipstick} = 1|\text{Young, Male}) &= \sigma(3.0 - 5.0 \cdot \text{Male} - 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Mouth Slightly Open} = 1|\text{Young, Smiling}) &= \sigma(-1.0 + 0.5 \cdot \text{Young} + \text{Smiling}) \\
\mathbb{P}(\text{Narrow Eyes} = 1|\text{Male, Young, Smiling}) &= \sigma(-0.5 + 0.3 \cdot \text{Male} + 0.2 \cdot \text{Young} + \text{Smiling}),
\end{aligned}
$$

where each variable either takes the value 0 or 1 and σ indicates the sigmoid. To generate data, we first simulate attributes from this binary Bayesian network, which we then pass as inputs to the GAN to simulate images (in addition to the random noise used by the GANs to simulate different images). In Figs. B.8 and B.9, we plot examples of the training images that were generated.

Predictive model We simulate a training set of 12,000 attribute-image pairs, and a validation set of 2,000 pairs. The training set is used to fit a classifier f, and the validation set is used for model selection. To build a classifier f, we use the ResNet-50 [He et al., 2016] model implemented in the python package `torch`. We add a final fully connected layer to adapt the ResNet model to a binary classification task, and fine-tune the model on the training data by (only) learning the weights and bias of the final layer. The model is trained using the negative log-likelihood criterion and an ADAM optimizer. The model is trained for 25 epochs and we select the model which after a full epoch had the best validation set performance. Given the learned model f, we simulate a separate

validation dataset of $n = 1{,}000$ samples, and make model predictions $f(X)$. We then compute the model accuracy as $\ell = \mathbf{1}\{f(X) = Y\}$, which is the input to computing the shift gradient and Hessian.

Estimation of shifted loss We apply the methods in Section 3.2 to estimate the worst-case shift to the distribution $\mathbb{P}$ (given by the binary probabilities above). For each conditional $\mathbb{P}(W_i | \operatorname{PA}(W_i))$, we consider a shift $\eta_{\delta_i}(\operatorname{PA}(W_i)) = \eta(\operatorname{PA}(W_i)) + \sum_{z \in \mathcal{Z}} \mathbf{1}\{\operatorname{PA}(W_i) = z\}\, \delta_i$, which corresponds to arbitrarily shifting the conditional distribution (see Appendix B.3.4). For example, for $W_i = \text{Bald}$, where $\eta(\text{Young}, \text{Male}) = -3.0 + 3.5 \cdot \text{Male} - 1.0 \cdot \text{Young}$, the shift would be

$$\eta_{\delta_{\text{Bald}}}(\text{Young}, \text{Male}) = \eta(\text{Young}, \text{Male}) + \begin{cases} \delta_{\text{Bald},0}, & \text{Young} = 0, \text{Male} = 0 \\ \delta_{\text{Bald},1}, & \text{Young} = 0, \text{Male} = 1 \\ \delta_{\text{Bald},2}, & \text{Young} = 1, \text{Male} = 0 \\ \delta_{\text{Bald},3}, & \text{Young} = 1, \text{Male} = 1. \end{cases} \tag{B.7}$$

For each W_i, this means that δ_i is $\mathbb{R}^{2^{|\operatorname{PA}(W_i)|}}$, and in total $\delta = (\delta_1, \ldots, \delta_8) \in \mathbb{R}^{31}$ (we do not consider shifts in the distribution of gender, since this is the label we are predicting).

We compute the shift gradient and Hessian using Theorem 1. In particular, since W_i is binary, the sufficient statistic is $T(W_i) = W_i$, so the shift gradients and Hessians given by Appendix B.3.4. See Appendix B.3.1 for a detailed walk through of computing the shift gradient and Hessian from a sample.

For any given δ, the shifted distribution of W_i is given by $\mathbb{P}_\delta(W_i = 1 | \operatorname{PA}(W_i)) = \sigma(\eta_{\delta_i})$, where η_{δ_i} is computed similar to (B.7), and σ is the sigmoid function. Then the importance sampling weights are given by

$$w_\delta = \prod_{i=1}^{8} \frac{\sigma(\eta_{\delta_i}(\operatorname{PA}(W_i)))}{\sigma(\eta(\operatorname{PA}(W_i)))}.$$

Using these, for any δ, we can estimate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ using (6) and (8), respectively.

B.5.2. Full table of worst-case shift in Section 4.2

In Section 4.2, we find the worst-case shift δ, and display the 5 largest components. In Table B.2, we display the full vector $\delta \in \mathbb{R}^{31}$, sorted by absolute value of the size of the component.

B.5.3. Sample images from training distribution in Section 4.2

In Fig. B.8, for the 4 attributes {Bald, Smiling, Wearing Lipstick, Male}, we display images generated from the training distribution $\mathbb{P}$ (i.e. by the GAN) with that particular attribute. In Fig. B.9 we show 10 randomly drawn images from the training distribution $\mathbb{P}$ as well as the test distribution $\mathbb{P}_\delta$ corresponding to the worst-case δ found in Section 4.2.

Conditional	δ_i
Bald — Male= 0, Young= 0	0.899
Bald — Male= 1, Young= 1	-0.800
Bald — Male= 1, Young= 0	-0.680
Wearing Lipstick — Male= 0, Young= 1	-0.618
Wearing Lipstick — Male= 0, Young= 0	-0.543
Eyeglasses — Young= 1	0.507
Mustache — Male= 1, Young= 0	-0.476
Mustache — Male= 0, Young= 0	0.449
Mustache — Male= 1, Young= 1	-0.415
Eyeglasses — Young= 0	0.399
Smiling — Male= 0, Young= 0	-0.261
Wearing Lipstick — Male= 1, Young= 0	0.205
Narrow Eyes — Male= 0, Smiling= 0, Young= 0	0.192
Mouth Slightly Open — Smiling= 1, Young= 1	0.191
Smiling — Male= 1, Young= 0	0.183
Narrow Eyes — Male= 1, Smiling= 1, Young= 1	0.179
Mouth Slightly Open — Smiling= 0, Young= 1	-0.153
Mustache — Male= 0, Young= 1	0.133
Bald — Male= 0, Young= 1	0.128
Mouth Slightly Open — Smiling= 1, Young= 0	-0.127
Narrow Eyes — Male= 0, Smiling= 1, Young= 0	-0.125
Wearing Lipstick — Male= 1, Young= 1	0.123
Narrow Eyes — Male= 1, Smiling= 1, Young= 0	-0.117
Narrow Eyes — Male= 0, Smiling= 0, Young= 1	0.106
Young — No parents	0.092
Narrow Eyes — Male= 0, Smiling= 1, Young= 1	0.057
Narrow Eyes — Male= 1, Smiling= 0, Young= 1	-0.050
Narrow Eyes — Male= 1, Smiling= 0, Young= 0	-0.039
Mouth Slightly Open — Smiling= 0, Young= 0	0.028
Smiling — Male= 1, Young= 1	0.028
Smiling — Male= 0, Young= 1	0.017

Table B.2.: Worst case shift in the $\delta \in \mathbb{R}^{31}$ identified by the Taylor approach in Section 4.2. Each entry corresponds to a shift in a conditional distribution given a particular outcome, and the squared sum of the entries equal $\lambda^2 = 4$.

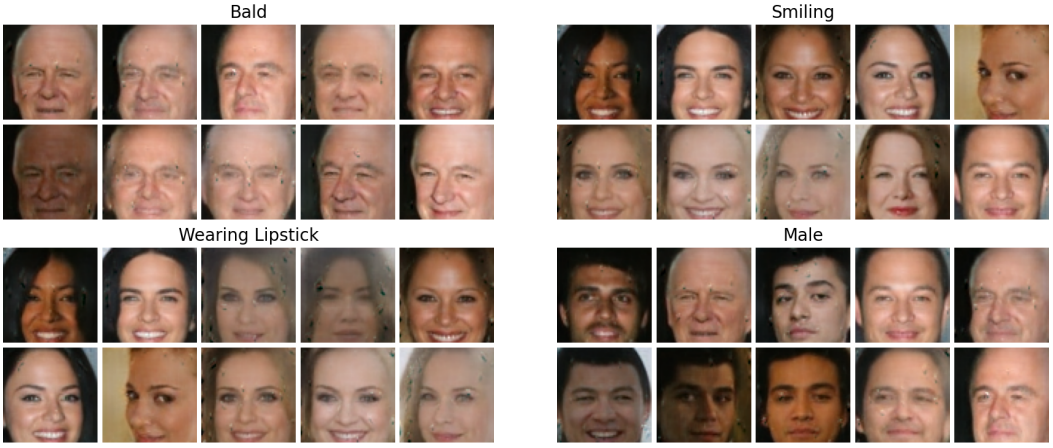


Figure B.8.: Examples of images from the training distribution $\mathbb{P}$. Each of the four groups (Bald, Smiling, Wearing Lipstick, Male) show training images who have that characteristic.

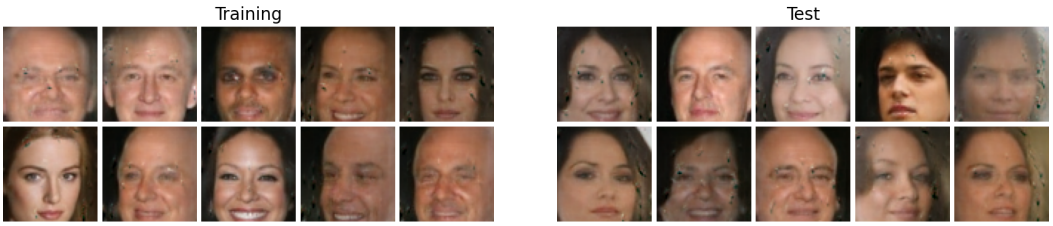


Figure B.9.: Examples of images from the training distribution $\mathbb{P}$ and the test distribution $\mathbb{P}_\delta$ that is characterized by the worst-case shift δ, see Fig. 4.

B.6. Proofs

B.6.1. Proof of Proposition 1

Proposition 1. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 4, $\operatorname{supp}(\mathbb{P}) = \operatorname{supp}(\mathbb{P}_\delta)$ and the density ratio $w_\delta := \mathbb{P}_\delta / \mathbb{P}$ is given by*

$$w_\delta(\mathbf{V}) = \exp\left(\sum_{i=1}^{m} s_i(Z_i;\delta_i)^\top T_i(W_i)\right) \exp\left(\sum_{i=1}^{m} h(\eta_i(Z_i)) - h(\eta(Z_i) + s_i(Z_i;\delta_i))\right).$$

Proof. By Definition 4 and Assumption 1, we have that

$$\mathbb{P}_\delta(\mathbf{V}) = \prod_{i=1}^{m} \mathbb{P}_{\delta_i}(W_i|Z_i) \prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)$$
$$\mathbb{P}(\mathbf{V}) = \prod_{i=1}^{m} \mathbb{P}(W_i|Z_i) \prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j).$$

It follows that the supports of $\mathbb{P}_\delta$ and $\mathbb{P}$ are the same: Since the exponential family density is given by the base measure $g_i(W_i)$ times a exponential term (which is always strictly positive), and since the terms $\prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)$ are shared between $\mathbb{P}_\delta$ and $\mathbb{P}$, their supports agree.

To get the density ratio, we take the ratio of $\mathbb{P}_\delta(\mathbf{V})$ and $\mathbb{P}(\mathbf{V})$, and the terms $V_j \in \mathbf{V} \setminus \mathbf{W}$ cancel:

$$\begin{aligned} w_\delta(\mathbf{V}) &= \frac{\mathbb{P}_\delta(\mathbf{V})}{\mathbb{P}(\mathbf{V})} \\ &= \prod_{i=1}^{m} \frac{\mathbb{P}_{\delta_i}(W_i|Z_i)}{\mathbb{P}(W_i|Z_i)}. \end{aligned}$$

By Definition 4 and Assumption 1, each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ is a δ_i-perturbation around the CEF distribuition $\mathbb{P}(W_i|Z_i)$, so plugging in the exponential family densities, we get

$$\begin{aligned} w_\delta(\mathbf{V}) &= \prod_{i=1}^{m} \frac{g(W_i) \exp\Big(\{\eta_i(Z_i) + s_i(Z_i;\delta_i)\}^\top T_i(W_i) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i))\Big)}{g(W_i) \exp\Big(\eta_i(Z_i)^\top T_i(W_i) - h_i(\eta_i(Z_i))\Big)} \\ &= \prod_{i=1}^{m} \exp\Big(s_i(Z_i;\delta_i) T_i(W_i) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i)) + h_i(\eta_i(Z_i))\Big) \\ &= \exp\Big(\sum_{i=1}^{m} s_i(Z_i;\delta_i) T_i(W_i)\Big) \exp\Big(\sum_{i=1}^{m} h_i(\eta_i(Z_i)) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i))\Big). \end{aligned}$$

□

B.6.2. Proof of Theorem 1

Theorem 1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 4, with intervened variables $\mathbf{W} = \{W_1, \dots, W_m\}$ and shift functions $s_i(Z_i;\delta_i)$, where $\delta = (\delta_1, \dots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \dots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, T_i(W_i) \middle| Z_i\right)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$*, where the* (i,j)*th block of size* $d_{\delta_i} \times d_{\delta_j}$ *equals*

$$\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, \varepsilon_{T_i|Z_i}\varepsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \varepsilon_{T|Z}\right] & i = j \\ \operatorname{cov}(\ell,\; D_{i,1}^\top \varepsilon_{T_i|Z_i}\varepsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla_{\delta_i}^k s_i(Z_i;\delta_i)|_{\delta=0}$*, is the gradient of the shift function for* $k = 1$*, and the Hessian for* $k = 2$*. Here,* $T_i(W_i)$ *is the sufficient statistic of* $\mathbb{P}(W_i|Z_i)$ *and* $\varepsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$*.*

Proof. For simplicity throughout, we use $h_i^{(1)}$ to denote the gradient of the log-partition function $\nabla h_i(\cdot)$ with respect to the arguments, which is a column vector of length d_{T_i}, and we use $h_i^{(2)}$ to denote the Hessian $\nabla^2 h_i(\cdot)$, which is a matrix of size $d_{T_i} \times d_{T_i}$. We also use $\eta_{\delta_i}(z_i)$ as short-hand for $\eta_i(z_i) + s_i(z_i;\delta_i)$.

Shift Gradient: By Definition 4, the probability density / mass function $\mathbb{P}_\delta$ factorizes as follows, where $\delta = (\delta_1, \ldots, \delta_m)$

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_i \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_i|\,\mathrm{PA}(V_i))\right), \tag{B.8}$$

and the gradient with respect to shift parameters δ_i is given by

$$\nabla_{\delta_i} p_\delta(v) = p_\delta(v)\nabla_{\delta_i} \log p_\delta(v) = p_\delta(v)\nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i)$$

where the last equality follows from additivity of the log-likelihood in the conditionals, the factorization above, and the fact that δ_i only enters into the given conditional distribution. Given the assumed form of $\log p_{\delta_i}(w_i|z_i)$ given in Definition 3, we can observe that

$$\begin{aligned} \nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i) &= \nabla_{\delta_i}\left[(\eta_i(z_i) + s_i(z_i;\delta_i))^\top T_i(w_i) - h_i(\eta(z_i) + s_i(z_i;\delta_i))\right] \\ &= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top T_i(w_i) - (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top \nabla h_i(\eta(z_i) + s_i(z_i;\delta_i)) \\ &= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \end{aligned} \tag{B.9}$$

where $\nabla_{\delta_i} s_i(z_i;\delta_i) \in \mathbb{R}^{d_{T_i} \times d_{\delta_i}}$, and $\nabla h_i(\eta(z_i) + s_i(z_i;\delta_i))$ is the gradient of the function $h_i : \mathbb{R}^{d_{T_i}} \to \mathbb{R}$, which is a column vector of length d_{T_i}. It follows from known properties of the log-partition function [Wainwright and Jordan, 2008, Proposition 3.1], that

$h_i^{(1)}(\eta_{\delta_i}(z_i)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$. This gives us that

$$\begin{aligned}\nabla_{\delta_i}\mathbb{E}_\delta[\ell] &= \mathbb{E}_\delta\left[\ell \cdot (\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i])\right] \\ &= \mathbb{E}_\delta\left[(\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top \mathbb{E}_\delta[\ell \cdot (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i])|Z_i]\right] \\ &= \mathbb{E}_\delta\left[(\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top \operatorname{cov}_\delta(\ell, T_i(W_i)|Z_i)\right],\end{aligned}$$

where the second equality follows from the tower property and Z_i-measurability of $\nabla_{\delta_i} s_i(Z_i;\delta_i)$, and the final equality follows from the definition of the conditional covariance. This expression, evaluated at $\delta = 0$, gives us the desired result, that

$$\mathrm{SG}_i^1 := \nabla_{\delta_i}\mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}(\ell, T_i(W_i)|Z_i)\right],$$

where $D_{i,1} = \nabla_{\delta_i} s_i(Z_i,\delta_i)|_{\delta=0}$. The result follows from the definition that gradients are taken entry-wise, giving $\mathrm{SG}^1 = (\mathrm{SG}_1^1, \ldots, \mathrm{SG}_m^1) \in \mathbb{R}^{d_{\delta_1}+\cdots d_{\delta_m}}$.

Shift Hessian (Diagonal): For the shift Hessian, we first compute the diagonal entries of $\nabla_\delta^2\mathbb{E}_\delta[\ell]|_{\delta=0}$, which are blocks of size $\mathbb{R}^{d_{\delta_i}\times d_{\delta_i}}$. We begin by computing the Hessian of the likelihood.

$$\begin{aligned}&\nabla_{\delta_i}^2 p_\delta(v) \\ &= \nabla_{\delta_i}\Big(p_\delta(v)\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\Big) \\ &= p_\delta(v)\Big((\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i))^{\otimes 2} + \nabla_{\delta_i}^2\log p_{\delta_i}(w_i|z_i)\Big) \\ &= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2}\{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\ &\qquad - \{\nabla_{\delta_i}^2 s_i(z_i;\delta_i)\}^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \\ &\qquad - \{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top h_i^{(2)}(\eta_{\delta_i}(z_i))\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}\Big), \\ &= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \Big(\big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2} - h_i^{(2)}(\eta_{\delta_i}(z_i))\Big)\{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\ &\qquad - \{\nabla_{\delta_i}^2 s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)\Big)\end{aligned}$$

where we use the notation $v^{\otimes 2} := vv^\top$, and we note that $\nabla_{\delta_i}^2 s(z_i;\delta_i)$ is a tensor of size $d_{T_i}\times d_{\delta_i}\times d_{\delta_i}$, and $\{\nabla_{\delta_i}^2 s_i(z_i;\delta_i)\}^\top h_i^{(1)}(\cdot)$ is a matrix of size $d_{\delta_i}\times d_{\delta_i}$, where the (m,n)'th entry is $\{\frac{\partial}{\partial\delta_{im}}\frac{\partial}{\partial\delta_{in}} s(z_i;\delta_i)\}^\top h^{(1)}(\cdot)$.

Now, using the fact that $h^{(1)}(\eta_{\delta_i}(z)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$ and $h^{(2)}(\eta_{\delta_i}(z_i)) = \operatorname{var}_\delta[T_i(W_i)|z_i]$ [Wainwright and Jordan, 2008, Proposition 3.1], and the definition $\varepsilon_{T_i|Z_i} = T_i(W_i) -$

$\mathbb{E}_\delta[T_i(W_i)|Z_i]$, we obtain

$$\begin{aligned}
&\nabla^2_{\delta_i}\mathbb{E}_\delta[\ell] \\
&= \mathbb{E}_\delta\left[\ell \cdot \{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}^\top \left(\varepsilon^{\otimes 2}_{T|Z_i} - \mathrm{var}_\delta(T_i(W_i)|Z_i)\right)\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}\right] \\
&\quad - \mathbb{E}_\delta\left[\ell \cdot \{\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\}^\top \varepsilon_{T_i|Z_i}\right] \\
&= \mathbb{E}_\delta\left[\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}^\top \mathrm{cov}_\delta\left(\ell, \varepsilon^{\otimes 2}_{T_i|Z_i}\middle| Z_i\right)\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}\right] \\
&\qquad - \mathbb{E}_\delta\left[\ell \cdot \{\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\}^\top \varepsilon_{T_i|Z_i}\right]
\end{aligned}$$

which gives the desired result when we evaluate at $\delta = 0$.

Shift Hessian (Off-Diagonal) For $i \neq j$, we have that

$$\begin{aligned}
&\nabla_{\delta_i}\nabla_{\delta_j} p_\delta(v) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= p_\delta(v)\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\big(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)\big)^\top \\
&= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))\Big) \\
&\qquad\qquad \Big(\{\nabla_{\delta_j} s_j(z_j;\delta_j)\}^\top (T_j(w_j) - h_j^{(1)}(\eta_{\delta_j}(z_j)))\Big)^\top
\end{aligned}$$

where the third line follows from the fact that $\nabla_{\delta_i}(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) = 0$, and the last line follows from the derivation of the gradient of the log-likelihood in (B.9). We can again use the fact that $h_i^{(1)}(\eta_{\delta_i}(Z_i)) = \mathbb{E}_\delta[T_i(W_i)|Z_i]$ and the shorthand $\varepsilon_{T_i|Z_i} \coloneqq T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]$ to write that

$$\begin{aligned}
&\nabla_{\delta_i}\nabla_{\delta_j}\mathbb{E}_\delta[\ell] \\
&= \mathbb{E}_\delta\Big[\ell \cdot \{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \Big((T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))\Big) \\
&\qquad\qquad \Big((T_j(w_j) - h_j^{(1)}(\eta_{\delta_j}(z_j)))\Big)^\top \{\nabla_{\delta_j} s_j(z_j;\delta_j)\}\Big]
\end{aligned}$$

and when we evaluate this expression at $\delta = 0$, we obtain

$$\nabla_{\delta_i}\nabla_{\delta_j}\mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[\ell \cdot D_{i,1}^\top \varepsilon_{T_i|Z_i}(\varepsilon_{T_j|Z_j})^\top D_{j,1}\right] = \mathrm{cov}(\ell, D_{i,1}^\top \varepsilon_{T_i|Z_i}\varepsilon_{T_j|Z_j}^\top D_{j,1}).$$

Where the last equality follows because $\mathbb{E}[D_{i,1}^\top \varepsilon_{T_i|Z_i}\varepsilon_{T_j|Z_j}^\top D_{j,i}] = 0$. To see this, note that one of W_i, W_j must be a non-descendant of the other, and we will assume without loss of generality that W_j is a non-descendant of W_i in the causal graph consistent with the

factorization given in (B.8), which implies that Z_j (the parents of W_j in the underlying graph) are also non-descendants of W_i.Thus, $W_i \perp\!\!\!\perp (W_j, Z_j)|Z_i$, because (W_j, Z_j) are both non-descendants of W_i. Then, observe that $D_{i,1}$ is a function of Z_i, and $\varepsilon_{T_i|Z_i}$ is a variable with zero-mean conditioned on Z_i. Thus, $\mathbb{E}[D_{i,1}^\top \varepsilon_{T_i|Z_i}|Z_i] = 0$, for all Z_i. Moreover, given Z_i, we have that $D_{i,1}^\top \varepsilon_{T_i|Z_i}$ is independent of $D_{j,1}^\top \varepsilon_{T_j|Z_j}$. As a result, we can write that

$$\begin{aligned}
\mathbb{E}[D_{i,1}^\top \varepsilon_{T_i|Z_i} \varepsilon_{T_j|Z_j}^\top D_{j,1}] &= \mathbb{E}[\mathbb{E}[D_{i,1}^\top \varepsilon_{T_i|Z_i} \varepsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= \mathbb{E}[\mathbb{E}[D_{i,1}^\top \varepsilon_{T_i|Z_i}|Z_i]\mathbb{E}[\varepsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= \mathbb{E}[0 \cdot \mathbb{E}[\varepsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= 0
\end{aligned}$$

□

B.6.3. Proof of Corollary 1

Corollary 1 (Simple shift in a single variable)**.** *Assume the setup of Theorem 1, restricted to a shift in a single variable W, and that $s(Z;\delta) = \delta$. Then $D_1 = 1$, $D_2 = 0$, and*

$$\mathrm{SG}^1 = \mathbb{E}\left[\mathrm{cov}\left(\ell, T(W)\middle| Z\right)\right] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}\left[\mathrm{cov}\left(\ell, \varepsilon_{T|Z}\varepsilon_{T|Z}^\top\middle| Z\right)\right],$$

where $T(W)$ is the sufficient statistic of W and $\varepsilon_{T|Z} := T(W) - \mathbb{E}[T(W)|Z]$.

Proof. We have $\nabla_\delta s(Z;\delta) = \nabla_\delta \delta = 1$ and $\nabla_\delta^2 s(Z;\delta) = \nabla_\delta^2 \delta = 0$. The result now follows from Theorem 1. □

B.6.4. Proof of Theorem 2

Theorem 2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem 1, with a shift in a single variable W, where $s(Z;\delta) = \delta$. Let $E_{\delta,\,Taylor}$ be the population Taylor estimate ((7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\,Taylor}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \sigma\left(\mathrm{cov}_{t\cdot\delta}(\ell, \varepsilon_{t\cdot\delta,T|Z}\varepsilon_{t\cdot\delta,T|Z}^\top) - \mathrm{cov}(\ell, \varepsilon_{0,T|Z}\varepsilon_{0,T|Z}^\top)\right) \cdot \|\delta\|^2,$$

where $T(W)$ is the sufficient statistic of $W|Z$ and $\varepsilon_{t\cdot\delta,T|Z} = T(W|Z) - \mathbb{E}_{t\cdot\delta}[T(W|Z)]$.

Proof. The expectation is continuous and twice-differentiable with respect to δ, because of the smoothness of the exponential family in the parameter, the fact that the shift function s is twice-differentiable, and because the support does not change. Thus, applying Taylors remainder theorem to the function $t \mapsto \mathbb{E}_{t\cdot\delta}[\ell]$, it follows that there exist

a $t_0 \in [0,1]$ such that

$$\mathbb{E}_{1\cdot\delta}[\ell] - \mathbb{E}_{0\cdot\delta}[\ell] - \left(\tfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \left(\tfrac{1}{2}\tfrac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0}. \tag{B.10}$$

We have $\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \mathrm{SG}^1$ and by the same arguments (see the proof of Theorem 1), it follows that $\left(\tfrac{1}{2}\tfrac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0} = \delta^\top \operatorname{cov}_{t_0\cdot\delta}(\ell, \varepsilon^{\otimes 2}_{t_0\cdot\delta,T|Z})\delta$. Plugging this in, and subtracting $\frac{1}{2}\delta^\top \mathrm{SG}^2\,\delta$ on both sides of (B.10) yields

$$\begin{aligned}\left|\mathbb{E}_\delta[\ell] - E_{\delta,\mathrm{Taylor}}\right| &= \tfrac{1}{2}\left|\delta^\top\left(\operatorname{cov}_{t_0\cdot\delta}(\ell, \varepsilon^{\otimes 2}_{t_0\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \varepsilon^{\otimes 2}_{0,T|Z})\right)\delta\right| \\ &\le \tfrac{1}{2}\sup_{t\in[0,1]}\left|\delta^\top\left(\operatorname{cov}_{t\cdot\delta}(\ell, \varepsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \varepsilon^{\otimes 2}_{0,T|Z})\right)\delta\right|.\end{aligned}$$

Let $K := \left(\operatorname{cov}_{t\cdot\delta}(\ell, \varepsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \varepsilon^{\otimes 2}_{0,T|Z})\right)$. Since K is symmetric and real valued, it is diagonalizeable, $K = U^\top \Lambda U$ for an orthonormal matrix U and diagonal matrix $\Lambda = \operatorname{diag}(\alpha_1, \ldots, \alpha_d)$. We then have

$$\begin{aligned}|\delta^\top K \delta| &= |\delta^\top U^\top \Lambda U \delta| \\ &= |(\Lambda^{1/2} U\delta)^\top (\Lambda^{1/2} U\delta)| \\ &= \|\Lambda^{1/2} U\delta\|_2^2 \\ &\le \|\Lambda^{1/2}\|_2^2 \|U\delta\|_2^2 \\ &= \sigma(K)\|\delta\|_2^2,\end{aligned}$$

where $\Lambda^{1/2} = \operatorname{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_d})$, $\|\cdot\|_2$ denotes the supremum-norm when applied to matrices and the 2-norm when applied to vectors and $\|U\delta\|_2 = \|\delta\|_2$ because $\|U\delta\|_2^2 = \delta^\top U^\top U\delta = \delta^\top\delta = \|\delta\|_2^2$, using orthonormality of U. Plugging in this inequality, we get that

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\mathrm{Taylor}}\right| \le \tfrac{1}{2}\sup_{t\in[0,1]} \sigma\left(\operatorname{cov}_{t\cdot\delta}(\ell, \varepsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \varepsilon^{\otimes 2}_{0,T|Z})\right)\|\delta\|_2^2,$$

which concludes the proof. □

B.6.5. Proof of Proposition B.1

Proposition B.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R} \cup \{+\infty, -\infty\}$. Let $p_+ := \mathbb{P}(\eta(Z) = +\infty)$, $p_- := \mathbb{P}(\eta(Z) = -\infty)$, and assume

that $p_+ + p_- < 1$. *Then, the marginal probability*

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ *whose range is* $(p_+, 1 - p_-)$,

Proof. Let F denote the event that $\eta(Z)$ is finite (i.e., $\eta(Z) \notin \{-\infty, +\infty\}$). Under F, the conditional probability function $\sigma(\eta(Z) + \delta)$ is a strictly monotonically increasing function of δ, and if $\eta(Z) \in \{-\infty, +\infty\}$, then the conditional probability is a constant function of δ (zero or one, respectively). Hence, we can write that

$$\mathbb{P}_\delta(W = 1) = \mathbb{P}_\delta(W = 1|F)(1 - p_+ - p_-) + p_+$$

and by assumption, $1 - p_+ - p_- > 0$. The marginal probability $\mathbb{P}_\delta(W = 1|F)$ is a strictly monotonically increasing function of δ, with a limit of 1 as $\delta \to \infty$, and a limit of 0 as $\delta \to -\infty$. As a result, it is bounded in $(p_+, 1 - p_-)$. □

B.6.6. Proof of Lemma B.1

Lemma B.1. *Suppose* $A \sim \mathcal{N}(\mu, \Sigma)$ *and that* (X, Y, H) *are generated according to* (B.2). *For* $\gamma \in \mathbb{R}^{d_X}$ *define* $\ell := (Y - \gamma^\top X)^2$. *Then there exist* $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ *such that for all shifts* $\delta \in \mathbb{R}^{d_A}$*:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top \delta,$$

where $\mathbb{E}_\delta$ *corresponds to taking the mean in the distribution where* $A \sim \mathcal{N}(\mu + \delta, \Sigma)$. *Further* $u_{\mu,\gamma} = 0$ *if* $\mu = 0$.

Proof. It follows from (B.2) that one can write $(X^\top, Y^\top, H^\top) = (1 - B)^{-1}(MA + \varepsilon)$, and for a given γ, there exist b_γ, κ_γ such that $Y - \gamma^\top X = b_\gamma^\top A + \kappa_\gamma^\top \varepsilon$ [Rothenhäusler et al., 2021]. In $\mathbb{P}_\delta$, we can write $A = \mu + \delta + \varepsilon_A$, where $\varepsilon_A \sim \mathcal{N}(0, \Sigma)$, for all values of μ and δ. Plugging this in yields

$$\begin{aligned}
\mathbb{E}_\delta[(Y - \gamma^\top X)^2] &= \mathbb{E}_\delta[(b_\gamma^\top A + \kappa_\gamma^\top \varepsilon)^2] \\
&= \mathbb{E}_\delta[(b_\gamma^\top(\mu + \delta + \varepsilon_A) + \kappa_\gamma^\top \varepsilon)^2] \\
&= \mathbb{E}[(b_\gamma^\top(\mu + \varepsilon_A) + \kappa^\top \varepsilon)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta \\
&= \mathbb{E}[(Y - \gamma^\top X)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta.
\end{aligned}$$

where we do not put a subscript on the expectation in the third line because it is taking expectations over ε_A and ε, both which do not depend on the choice of μ and δ. The statement of the lemma follows by letting $u_{\mu,\gamma} = 2b_\gamma^\top \mu$ and $v_\gamma = \sqrt{2} b_\gamma$. □

B.6.7. Proof of Proposition B.2

Proposition B.2. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to (B.2). Then the shift gradient and Hessian are given by*

$$\mathrm{SG}^1 = \mathrm{cov}(\ell, \Sigma^{-1}A) \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathrm{cov}(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top})$$

and the loss under a mean shift of δ in A is given by

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta,$$

where $\ell := (Y - \gamma^\top X)^2$ and $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$.

Proof. Similar to Lemma B.1, we rewrite $Y - \gamma^\top X = b_\gamma^\top A + \kappa^\top \varepsilon$, and by rewriting $A = \mu + \delta + \varepsilon_A$, where $\varepsilon_A \sim \mathcal{N}(0, \Sigma)$, we obtain

$$\begin{aligned}
\mathbb{E}_\delta[(Y - \gamma^\top X)^2] = \mathbb{E}(b_\gamma^\top(\mu + \varepsilon_A) + \kappa^\top \varepsilon)^2 && \text{(B.11)}\\
+ (2b_\gamma^\top \mu)\delta^\top b && \text{(B.12)}\\
+ \delta^\top b b^\top \delta. && \text{(B.13)}
\end{aligned}$$

We recognize that (B.11) equals $\mathbb{E}(Y - \gamma^\top X)^2$. Similarly, we now show that (B.12) and (B.13) match the shift gradients (multiplied appropriately with δ).

First, we assume that $\Sigma = \mathrm{Id}$. Since A is a Gaussian with (known) mean Id, the sufficient statistic is $T(A) = A$. Hence, according to Theorem 1, we can compute the shift gradient as

$$\mathrm{SG}^1 = \mathrm{cov}(A, \ell) = \mathrm{cov}(A, (Y - \gamma^\top X)^2) = \mathrm{cov}(A, (b_\gamma^\top A)^2).$$

We can calculate the i'th entrance of this vector as:

$$\begin{aligned}
\mathrm{SG}^1 = \mathrm{cov}(A_i, (b_\gamma^\top A)^2) &= \mathrm{cov}(A_i - \mu_i, (b_\gamma^\top A)^2))\\
&= \mathrm{cov}(A_i - \mu_i, b_{\gamma,i}^2 A_i^2 + 2\sum_{j \neq i} b_i b_j A_i A_j)\\
&= b_{\gamma,i}^2\, \mathrm{cov}(A_i - \mu_i, A_i^2) + 2b_{\gamma,i} \sum_{j \neq i} b_j\, \mathrm{cov}(A_i - \mu_i, A_i A_j),
\end{aligned}$$

where in the first equality we use that subtracting a constant doesn't change the covariance, and we use independence of A_i from $A_j A_{j'}$ when $i \notin \{j, j'\}$. Using the assumption that A_i has unit variance, we now get that

$$\begin{aligned}
\mathrm{cov}(A_i - \mu_i, A_i^2) &= \mathbb{E}[A_i^3 - \mu_i A_i^2] = (\mu_i^3 + 3\mu_i) - \mu_i(\mu_i^2 + 1) = 2\mu_i\\
\mathrm{cov}(A_i - \mu_i, A_i A_j) &= \mathbb{E}[A_i^2 - A_i\mu_i]\mathbb{E}[A_j] = (\mu_i^2 + 1 - \mu_i^2)\mu_j = \mu_j.
\end{aligned}$$

By plugging in, we obtain

$$\begin{aligned}\mathrm{SG}^1(\mu_i) &= 2b_{\gamma,i}^2\mu_i + 2b_{\gamma,i}\sum_{j\neq i} b_j\mu_j \\ &= 2b_{\gamma,i}b_\gamma^\top\mu.\end{aligned}$$

Since this was element-wise, we obtain that the full vector is $\mathrm{SG}^1 = 2b_\gamma b_\gamma^\top\mu$, which, when multiplied with δ yields (B.12).

We compute SG^2 similarly. The diagonal entries are given by

$$\begin{aligned}\mathrm{SG}_{i,i}^2 &= \mathrm{cov}((A_i-\mu_i)^2, (b_\gamma^\top A)^2) \\ &= \mathrm{cov}((A_i-\mu_i)^2, b_{\gamma,i}^2A_i^2 + b_{\gamma,i}\sum_{j\neq i} b_{\gamma,j}A_iA_j) \\ &= b_{\gamma,i}^2\,\mathrm{cov}((A_i-\mu_i)^2, A_i^2) + b_{\gamma,i}\sum_{j\neq i} b_{\gamma,j}\,\mathrm{cov}((A_i-\mu_i)^2, A_iA_j).\end{aligned}$$

Because $\Sigma = \mathrm{Id}$, the second through fourth moments of A_i are given by $\mathbb{E}[A_i^2] = \mu_i^2 + 1$, $\mathbb{E}[A_i^3] = \mu_i^3 + 3\mu_i$ and $\mathbb{E}[A_i^4] = \mu_i^4 + 6\mu_i^2 + 3$. Using this, we get

$$\begin{aligned}\mathrm{cov}((A_i-\mu_i)^2, A_i^2) &= \mathbb{E}[A_i^4 - 2\mu_iA_i^3 + \mu_i^2A_i^2] - \mathbb{E}[(A_i-\mu_i)^2]\mathbb{E}[A_i^2] \\ &= (\mu_i^4 + 6\mu_i^2 + 3) - 2\mu_i(\mu_i^3 + 3\mu_i) + \mu_i^2(\mu_i^2+1) - 1\cdot(\mu_i^2+1) \\ &= 2,\end{aligned}$$

and for $j \neq i$:

$$\begin{aligned}\mathrm{cov}((A_i-\mu_i)^2, A_iA_j) &= \mathrm{cov}((A_i-\mu_i)^2, (A_i-\mu_i)A_j) + \mathrm{cov}((A_i-\mu_i)^2, \mu_iA_j) \\ &= \mathrm{cov}((A_i-\mu_i)^2, (A_i-\mu_i)A_j) \\ &= \mathbb{E}[(A_i-\mu_i)^3]\mathbb{E}[A_j] - \mathbb{E}[(A_i-\mu_i)^2]\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_j] \\ &= 0 - 0,\end{aligned}$$

using linearity of the covariance, that $A_i \perp\!\!\!\perp A_j$ and that the first and third moments are zero for a centered Gaussian $A_i - \mu_i$. Plugging this in, we get that the diagonal entries are given by

$$\mathrm{SG}_{i,i}^2 = 2b_{\gamma,i}^2.$$

We can compute the off-diagonal entries similarly. For $i \neq j$, we have:

$$\begin{aligned}\mathrm{SG}_{i,j}^2 = \mathrm{cov}\Big(&(A_i-\mu_i)(A_j-\mu_j), \\ &b_{\gamma,i}^2A_i^2 + b_{\gamma,j}^2A_j^2 + 2b_{\gamma,i}b_{\gamma,j}A_iA_j + 2\sum_{v\notin\{i,j\}} b_{\gamma,i}b_{\gamma,v}A_iA_v + b_{\gamma,j}b_{\gamma,v}A_jA_v\Big).\end{aligned} \tag{B.14}$$

Using the independence of A_i and A_j, we have

$$\begin{aligned}&\operatorname{cov}((A_i-\mu_i)(A_j-\mu_j),A_i^2)\\&=\mathbb{E}[A_i^2(A_i-\mu_i)]\underbrace{\mathbb{E}[A_j-\mu_j]}_{=0}-\underbrace{\mathbb{E}[A_i-\mu_i]}_{=0}\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_i^2]\\&=0,\end{aligned}$$

and similarly $\operatorname{cov}((A_i-\mu_i)(A_j-\mu_j),A_j^2)=0$. Using the same reasoning, for $v\notin\{i,j\}$

$$\begin{aligned}&\operatorname{cov}((A_i-\mu_i)(A_j-\mu_j),A_iA_v)\\&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v]-\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v]\\&=0,\end{aligned}$$

and the same for $\operatorname{cov}((A_i-\mu_i)(A_j-\mu_j),A_jA_v)$. Finally, we have

$$\begin{aligned}&\operatorname{cov}((A_i-\mu_i)(A_j-\mu_j),A_iA_j)\\&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[(A_j-\mu_j)A_j]-\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_i]\mathbb{E}[(A_j-\mu_j)]\mathbb{E}[A_j]\\&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[(A_j-\mu_j)A_j]\\&=\mathbb{E}[A_i^2-\mu_iA_i]\mathbb{E}[A_j^2-\mu_jA_j]\\&=[(\mu_i^2+1)-\mu_i^2][[(\mu_j^2+1)-\mu_j^2]]\\&=1.\end{aligned}$$

Plugging into (B.14), we get that

$$\mathrm{SG}_{i,j}^2=2b_{\gamma,i}b_{\gamma,j},$$

and hence for both diagonal and off-diagonal entries, $\mathrm{SG}_{i,j}^2=2b_{\gamma,i}b_{\gamma,j}$, implying that

$$\mathrm{SG}^2=2b_\gamma b_\gamma^\top.$$

In particular $\frac{1}{2}\delta^\top\,\mathrm{SG}^2\,\delta$ matches (B.13).

Finally, we consider the case $\Sigma\neq\mathrm{Id}$. Let $\Sigma^{-1/2}$ be the 'square-root' of Σ^{-1}, such that $\Sigma^{-1/2}\Sigma^{-\top/2}$ (where the latter denotes $(\Sigma^{-1/2})^\top$.[5]

The sufficient statistics for the mean in a multivariate Gaussian distribution with known variance is given by $T(A)=\Sigma^{-1}A$. We then have

$$\begin{aligned}\mathrm{SG}^1&=\operatorname{cov}(\Sigma^{-1}A,(b_\gamma^\top A)^2)\\&=\Sigma^{-1/2}\operatorname{cov}(\Sigma^{-1/2}A,((\Sigma^{1/2}b_\gamma)^\top\Sigma^{-1/2}A)^2)\\&=\Sigma^{-1/2}\operatorname{cov}_{\tilde\mu}(\tilde A,(\tilde b_\gamma^\top\tilde A)^2),\end{aligned}$$

where $\tilde A=\Sigma^{-1/2}A=\sim\mathcal{N}(\tilde\mu,\mathrm{Id})$, $\tilde\mu=\Sigma^{-1/2}\mu$ and $\tilde b_\gamma=\Sigma^{1/2}b_\gamma$. In particular, since $\tilde A$

[5]Formally, if $\Sigma^{-1}=U\Lambda U^\top$ where $\Lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_{d_A})$, define $\Sigma^{-1/2}:=U\operatorname{diag}(\sqrt{\lambda_1},\dots,\sqrt{\lambda_{d_A}})$.

has unit variance, we can use the above derivations to obtain

$$\mathrm{SG}^1 = 2\Sigma^{-1/2}(\tilde{b}_\gamma \tilde{b}_\gamma^\top \tilde{\mu}) = 2b_\gamma b_\gamma^\top \mu.$$

In particular, the first shift gradient is the when $\Sigma \neq \mathrm{Id}$ as when $\Sigma = \mathrm{Id}$. Similarly,

$$\begin{aligned}
\mathrm{SG}^2 &= \mathrm{cov}(\Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top}, (b_\gamma^\top A)^2) \\
&= \mathrm{cov}(\Sigma^{-1/2}\Sigma^{-1/2}(A-\mu)(A-\mu)^\top \Sigma^{-\top/2}\Sigma^{-\top/2}, (\Sigma^{1/2} b_\gamma)^\top \Sigma^{-1/2} A)^2) \\
&= \Sigma^{-1/2}\,\mathrm{cov}_{\tilde{\mu}}((\tilde{A}-\tilde{\mu})(\tilde{A}-\tilde{\mu})^\top, (\tilde{b}_\gamma^\top \tilde{A})^2)\Sigma^{-\top/2} \\
&= \Sigma^{-1/2} 2\tilde{b}_\gamma \tilde{b}_\gamma^\top \Sigma^{-\top/2} \\
&= 2b_\gamma b_\gamma^\top.
\end{aligned}$$

Hence, also when $\Sigma \neq \mathrm{Id}$, the terms of (B.12) and (B.13) matches the expression given by SG^1 and SG^2. This concludes the proof. □

C. Appendix to Regularizing towards Causal Invariance: Linear Models with Proxies

Supplementary Materials

The supplementary materials are organized as follows

- (Appendix C.1): First, we give a simple 1D example to build intuition for the theoretical results.
- (Appendix C.2): In the context of Section 3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W := \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Fig. C.2).
- (Appendix C.3): Proofs for results stated in the main paper.
- (Appendix C.4): Additional results (and proofs) for Proxy Targeted Anchor Regression (PTAR) and Cross-Proxy TAR, deferred from the main paper.
- (Appendix C.5): Details for implementation of all experiments
- (Appendix C.6): Additional synthetic experimental results

C.1. An example for building intuition

To illustrate the problem, consider the following setup, where we observe A, X, Y at training time, and wish to learn a predictor $\hat{y} = \alpha + \gamma x$ that will generalize to a new environment where $\mathbb{P}_{te}(A) \neq \mathbb{P}_{tr}(A)$.

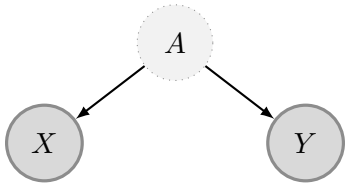


Figure C.1.: Simple example where $X, Y, A \in \mathbb{R}$.

Suppose that our data is generated under $\mathbb{P}_{tr}$ as follows

$$\begin{aligned} A &= \varepsilon_A, & \varepsilon_A &\sim \mathcal{N}(0,1) \\ X &= A + \varepsilon_X, & \varepsilon_X &\sim \mathcal{N}(0,\sigma_X^2) \\ Y &= A + \varepsilon_Y, & \varepsilon_Y &\sim \mathcal{N}(0,\sigma_Y^2), \end{aligned}$$

where $\varepsilon_A, \varepsilon_X, \varepsilon_Y$ are jointly independent. This simple example demonstrates a few concepts:

- Assuming $\sigma_X^2 > 0$, the conditional expectation $\mathbb{E}[Y \mid X]$ changes as the distribution of A changes.
- We can write the residuals $Y - \hat{Y}$ as a linear function in A and the noise variables. This holds, even if the errors are non-Gaussian.
- The test population MSE is a convex function of α, γ.

In particular, we will see that the parameters α, γ trade off between the variance of A and ε_X: There exists an invariant solution, where $\alpha = 0, \gamma^* = 1$, such that the MSE is completely independent of A, but this is only optimal in the setting where $\text{var}(A) \to \infty$.

Conditional Expectation depends on A Starting with the assumption that A, X, Y are multivariate Gaussian, we can write down the optimal predictor in the target environment, supposing that at test time $\mathbb{P}_{te}(A) \overset{(\text{d})}{=} \mathcal{N}(\mu_A, \sigma_A^2)$.

$$\begin{aligned} \mathbb{E}_{te}[Y \mid X = x] &= \mathbb{E}_{te}[Y] + \frac{\text{cov}_{te}(X,Y)}{\text{var}_{te}(X)} \cdot (x - \mathbb{E}_{te}[X]) \\ &= \mu_A + \underbrace{\frac{\sigma_A^2}{\sigma_A^2 + \sigma_X^2}}_{\gamma} \cdot (x - \mu_A) \\ &= \mu_A(1-\gamma) + \gamma x, \end{aligned}$$

where if $\varepsilon_X = 0$, then $\gamma = 1$ and the optimal solution does not depend on the parameters of A, and is given by

$$\mathbb{E}_{te}[Y|X = x] = x. \tag{C.1}$$

However, for any $\sigma_x^2 > 0$, the optimal solution under $\mathbb{P}_{te}(A)$ depends on μ_A, σ_A^2.

Rewriting residuals Regardless of whether the Gaussian assumption holds, for a given predictor $\hat{Y} = \alpha + \gamma x$, we can write the error $Y - \hat{Y}$ as a function that is linear in A and the noise variables

$$\begin{aligned} Y - \hat{Y} &= (A + \varepsilon_Y) - \gamma(A + \varepsilon_X) - \alpha \\ &= A(1-\gamma) + (\varepsilon_Y - \gamma\varepsilon_X - \alpha). \end{aligned}$$

Optimizing for a known target distribution The mean squared error $\mathbb{E}[(Y - \hat{Y})^2]$ can be written as a function of α, γ, and the mean and variance of A under $\mathbb{P}_{te}(A)$. Here, all expectations are taken with respect to the test distribution.

$$\begin{aligned}\mathbb{E}_{te}[(y - \hat{y})^2] &= \mathbb{E}_{te}[\mathbb{E}_{te}[(y - \hat{y})^2 \mid A]] \\ &= \alpha^2 - 2\alpha\mathbb{E}_{te}[A](1 - \gamma) \\ &\quad + (1 - \gamma)^2\mathbb{E}_{te}[A^2] + \gamma^2\sigma_x^2 + \sigma_y^2. \end{aligned} \tag{C.2}$$

By first-order conditions, this expression is minimized by

$$\alpha^* = \mu_A(1 - \gamma^*) \qquad \gamma^* = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_X^2}. \tag{C.3}$$

When $\sigma_A^2 \to \infty$, then $\gamma^* \to 1$ from (C.3). This is intuitive, because in (C.2), $\gamma = 1$ renders the MSE functionally independent of the distribution of A.

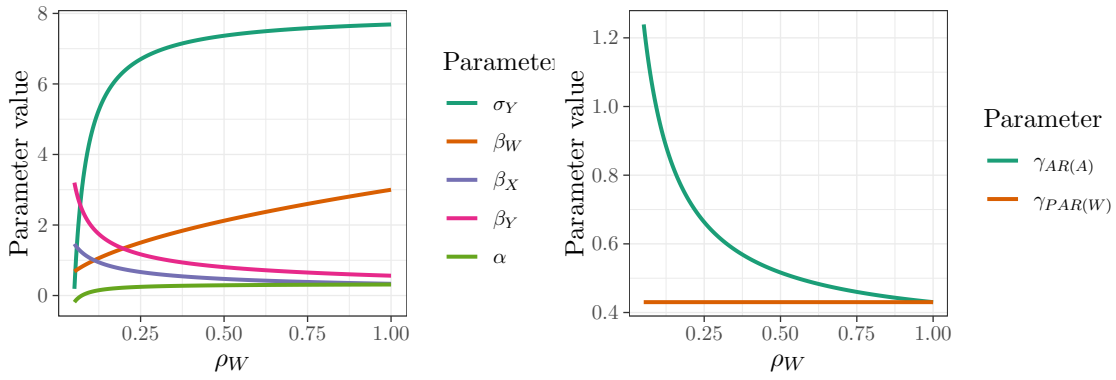
Optimizing for a worst-case distribution (C.3) shows the optimal solution under a known target distribution, if μ_A, σ_A^2 were known in advance. However, a similar intuition applies to the case where $\mathbb{P}_{te}(A)$ is unknown, but we expect it to lie in a particular class. Consider interventions of the form $do(A := \nu)$, where we constrain ν to lie in the set of random variables $C(\lambda) := \{\nu : \mathbb{E}[\nu^2] \leq \lambda\}$. In this case, our worst-case loss is given by

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y - \hat{Y})^2] \\ &= \sup_{\nu \in C(\lambda)} (1 - \gamma)\left[-2\alpha\mathbb{E}[\nu] + (1 - \gamma)\mathbb{E}[\nu^2]\right] \\ &\quad + \alpha^2 + \gamma^2\sigma_X^2 + \sigma_Y^2,\end{aligned}$$

where the last line does not depend on ν. We observe that $\alpha^* = 0$, by analyzing two cases. First, if $\gamma = 1$, then the first term is eliminated, and the only term that depends on α is α^2. Second, if $\gamma \neq 1$, then $(1 - \gamma)^2 > 0$, the first term is partially maximized when $\mathbb{E}[\nu^2] = \lambda$, and if $\alpha \neq 0$, then the expression can be made even larger by choosing a deterministic $\nu = \pm\sqrt{\lambda}$ (instead of e.g., a random $\nu \sim \mathcal{N}(0, \lambda^2)$), depending on the sign of $\alpha(1-\gamma)$. From this (and the presence of the α^2 term in the second line) it follows that $\alpha^* = 0$, in this case as well. When $\alpha = 0$, the supremum is obtained by any random or deterministic ν such that $\mathbb{E}[\nu^2] = \lambda$.

With $\alpha^* = 0$ and taking $\mathbb{E}[\nu^2] = \lambda$ in the supremum, this expression simplifies to

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y - \hat{Y})^2] \\ &= (1 - \gamma)^2\lambda + \gamma^2\sigma_X^2 + \sigma_Y^2.\end{aligned}$$



(a) (b)

Figure C.2.: (a) SCM parameters that all give rise to the same observational distribution, and observe that (b) the parameter $\gamma_{AR(A)}$ (as if A were observed) can diverge substantially from the solution $\gamma_{PAR(W)}$, when a single proxy is available. $\lambda = 5$ for this example.

Differentiating with respect to γ, we obtain

$$\gamma^* = \frac{\lambda}{\sigma_X^2 + \lambda}.$$

Here, λ trades off accuracy and stability; As $\lambda \to \infty$, we recover the solution where $\gamma^* = 1$, but for situations where σ_X^2 is large and λ is bounded, we are better off choosing $\gamma^* < 1$.

C.2. Example: Non-identifiability of Ω_W

Overview In the context of Section 3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W := \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Fig. C.2).

Setup If $(X, Y, W) \in \mathbb{R}^3$ is distributed multivariate normal with zero mean, then their covariance matrix fully determines the observed distribution. Let that covariance matrix

be denoted by $\Sigma_{(X,Y,W)} \in \mathbb{R}^{3\times 3}$, which gives us six observed moments of the distribution

$$\Sigma_{(X,Y,W)} \coloneqq \begin{pmatrix} \mathbb{E}[X^2] & \cdot & \cdot \\ \mathbb{E}[XY] & \mathbb{E}[Y^2] & \cdot \\ \mathbb{E}[WX] & \mathbb{E}[WY] & \mathbb{E}[W^2] \end{pmatrix},$$

where we only show the lower triangular portion, since the matrix is symmetric. Suppose that we knew that this observed distribution was generated by the following SCM, but that we do not know the values for the parameters $(\beta_W, \beta_X, \beta_Y, \alpha, \sigma_W^2, \sigma_X^2, \sigma_Y^2)$

$$\begin{aligned} A &\coloneqq \varepsilon_A & \varepsilon_A &\sim \mathcal{N}(0,1) \\ W &\coloneqq \beta_W A + \varepsilon_W & \varepsilon_W &\sim \mathcal{N}(0,\sigma_W^2) \\ X &\coloneqq \beta_X A + \varepsilon_X & \varepsilon_X &\sim \mathcal{N}(0,\sigma_X^2) \\ Y &\coloneqq \alpha X + \beta_Y A + \varepsilon_Y & \varepsilon_Y &\sim \mathcal{N}(0,\sigma_Y^2), \end{aligned}$$

where $\varepsilon_A, \varepsilon_W, \varepsilon_X, \varepsilon_Y$ are jointly independent. We can attempt to identify the parameters using the following relationships implied by the SCM, and matching these to the moments that we observe

$$\begin{aligned} \mathbb{E}[WX] &= \beta_W \beta_X \\ \mathbb{E}[XY] &= \beta_Y \beta_X + \alpha \mathbb{E}[X^2] \\ \mathbb{E}[WY] &= \beta_W(\beta_Y + \alpha \beta_X) \\ \mathbb{E}[W^2] &= \beta_W^2 + \sigma_W^2 \\ \mathbb{E}[X^2] &= \beta_X^2 + \sigma_X^2 \\ \mathbb{E}[Y^2] &= \alpha^2 \mathbb{E}[X^2] + 2\alpha\beta_Y\beta_X + \beta_Y^2 + \sigma_Y^2 \end{aligned}$$

However, as we will see, this does not identify the parameters. In particular, there is a set of parameterizations which all give rise to the same observed distribution, and which imply different values of the signal-to-variance ratio $\rho_W \coloneqq \beta_W^2/(\beta_W^2 + \sigma_W^2)$.

A class of observationally equivalent SCMs Let $\theta \coloneqq (\beta_W, \beta_X, \beta_Y, \alpha, \sigma_W^2, \sigma_X^2, \sigma_Y^2) \in \mathbb{R}^7$ be the parameters of the SCM, and let $\Sigma = f(\theta)$ be the covariance matrix over (X, Y, W) implied by these parameters.

For any covariance matrix Σ, there exists a subset $C \subset [0,1]$ such that for any $\rho_W \in C$, we can write the parameters as a function of ρ_W, such that $f(\theta(\rho_W)) = \Sigma$. The set C is constrained by the observed moments: In particular, as we show below, $\rho_W \geq \operatorname{corr}(W,X)^2$ due to the constraint that $\sigma_X^2 \geq 0$, and the condition that $\sigma_Y^2 \geq 0$ also imposes a lower bound. In particular, for the covariance matrix below, we demonstrate

numerically that $[0.06, 1] \subset C$.

$$\Sigma_{(X,Y,W)} \coloneqq \begin{pmatrix} 9 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 9 \end{pmatrix}.$$

We now give a strategy for constructing $\theta(\rho_W)$, given a desired ρ_W (including checking the constraint that this $\rho_W \in C$). Suppose that W and X are positively correlated, as in this example. Fixing some $\rho_W \in [0, 1]$, we start by writing β_W, σ_W as functions of ρ_W, where

$$\begin{aligned} \beta_W &\coloneqq \sqrt{\mathbb{E}[W^2]\rho_W} \\ \sigma_W^2 &\coloneqq \mathbb{E}[W^2](1 - \rho_W). \end{aligned}$$

The first constraint, that $\sigma_X^2 \geq 0$, can be captured as follows. Let $\rho_X \coloneqq \beta_X^2/\mathbb{E}[X^2]$. Observe that $\sqrt{\rho_X \rho_W} = \text{corr}(W, X)$. This implies a lower bound on ρ_W, given by $\rho_W \geq \text{corr}(W, X)^2$, since $\rho_X \leq 1$ due to $\sigma_X^2 \geq 0$. This also implies that ρ_X is determined uniquely by ρ_W, and is given by $\rho_X = \text{corr}(W, X)^2/\rho_W$. From this we can write

$$\begin{aligned} \beta_X &\coloneqq \sqrt{\mathbb{E}[X^2]\rho_X} \\ \sigma_X^2 &\coloneqq \mathbb{E}[X^2](1 - \rho_X). \end{aligned}$$

These choices for $(\beta_W, \sigma_W^2, \beta_X, \sigma_X^2)$ match the observed moments $\mathbb{E}[X^2], \mathbb{E}[W^2], \mathbb{E}[WX]$. Then the rest of the parameters can be found as follows, where β_W, β_X are fixed as above

$$\begin{aligned} \beta_Y &\coloneqq \frac{1}{\beta_W(1 - \rho_X)} \left(\mathbb{E}[WY] - \frac{\mathbb{E}[XY]\mathbb{E}[WX]}{\mathbb{E}[X^2]} \right) \\ \alpha &\coloneqq \frac{\mathbb{E}[XY] - \beta_Y \beta_X}{\mathbb{E}[X^2]} \\ \sigma_Y^2 &\coloneqq \mathbb{E}[Y^2] - \beta_Y^2 - 2\alpha\beta_Y\beta_X - \alpha^2\mathbb{E}[X^2] \end{aligned}$$

where all of these are functions of ρ_W, in that β_W, β_X are functions of ρ_W. It remains to verify that for a given choice of ρ_W, we satisfy the constraint that $\sigma_Y^2 \geq 0$. For simplicity, we check this constraint computationally in the context of Example 1, for a range of values of ρ_W, and we give the set of observationally-equivalent parameters in Fig. C.2a, where valid values of ρ_W range over $[0.06, 1]$.

Next we show that the Proxy Anchor Regression estimator, $\gamma_{PAR(W)}$, differs from the Anchor Regression estimator, $\gamma_{AR(A)}$, and more so when ρ_W becomes small. This is shown in Fig. C.2b, for $\lambda = 5$, and we give the relevant computations here.

Solution to PAR(W) If we have a single proxy, then we can write down the optimization problem (10) as

$$\begin{aligned}
&\min_{\gamma} \mathbb{E}[(Y-\gamma X)^2] + \lambda \mathbb{E}[(Y-\gamma X)W]^2 \mathbb{E}[W^2]^{-1} \\
=&\min_{\gamma} \mathbb{E}[Y^2] - 2\gamma \mathbb{E}[YX] + \gamma^2 E[X^2] \\
&+ \lambda(\mathbb{E}[YW] - \gamma \mathbb{E}[XW])^2 \mathbb{E}[W^2]^{-1},
\end{aligned}$$

from which we obtain the optimal solution

$$\gamma_{PAR(W)} = \frac{\mathbb{E}[YX]\mathbb{E}[W^2] + \lambda \mathbb{E}[YW]}{\mathbb{E}[X^2]\mathbb{E}[W^2] + \lambda \mathbb{E}[XW]}.$$

Solution to AR(A) First, we can write the residual as

$$\begin{aligned}
Y - \hat{Y} &= Y - \gamma X \\
&= \alpha X + \beta_Y A + \varepsilon_Y - \gamma \beta_X A - \gamma \varepsilon_X \\
&= \alpha(\beta_X A + \varepsilon_X) + \beta_Y A + \varepsilon_Y - \gamma \beta_X A - \gamma \varepsilon_X \\
&= A((\alpha - \gamma)\beta_X + \beta_Y) + (\alpha - \gamma)\varepsilon_X + \varepsilon_Y,
\end{aligned}$$

such that the expected squared error is given by

$$\begin{aligned}
&\mathbb{E}_{do(A:=\nu)}(Y - \hat{Y})^2 \\
&= ((\alpha - \gamma)\beta_X + \beta_Y)^2 \mathbb{E}[\nu^2] + (\alpha - \gamma)^2 \sigma_X^2 + \sigma_Y^2,
\end{aligned} \tag{C.4}$$

and when $\nu \in \{\nu : \mathbb{E}[\nu^2] \leq (1+\lambda)\}$, taking the supremum involves replacing $\mathbb{E}[\nu^2]$ with $(1+\lambda)$. Optimizing (C.4) with respect to γ, we obtain

$$\begin{aligned}
&\frac{\partial}{\partial \gamma}\left[((\alpha - \gamma)\beta_X + \beta_Y)^2(1+\lambda) + (\alpha - \gamma)^2\sigma_X^2 + \sigma_Y^2\right] \\
&= -2\beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1+\lambda) - 2(\alpha - \gamma)\sigma_X^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
0 &= \beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1+\lambda) + (\alpha - \gamma)\sigma_X^2 \\
&= (\alpha\beta_X^2 + \beta_X\beta_Y)(1+\lambda) - \gamma\beta_X^2(1+\lambda) + \alpha\sigma_X^2 - \gamma\sigma_X^2,
\end{aligned}$$

so that the optimal choice of γ is given by

$$\gamma_{AR(A)} = \frac{(\alpha\beta_X^2 + \beta_X\beta_Y)(1+\lambda) + \alpha\sigma_X^2}{\beta_X^2(1+\lambda) + \sigma_X^2}.$$

If $\lambda = -1$, this recovers the causal effect of X on Y, and if $\lambda \to \infty$, this recovers a set of coefficients that are invariant to variation in A, as can be seen by plugging the resulting

coefficient $\gamma = \alpha + \beta_Y/\beta_X$ into (C.4).

C.3. Proofs

C.3.1. Auxiliary results

First, we show that the proof of Theorem 1 of Rothenhäusler et al. [2021] can be decomposed into two parts, and use this observation to simplify the proof of our Theorem 1. Proposition C.1 establishes that ℓ_{PLS} can be written as a quadratic form in the structural parameters $w_\gamma^\top M_A$. Proposition C.2 is a straightforward generalization of the techniques used in Rothenhäusler et al. [2021], and establishes that any regularization term that can be written in this way naturally implies a robustness guarantee.

By Assumption 1, our SCM can be written in the following form, where $\varepsilon \perp\!\!\!\perp A$, and all variables are mean-zero and have bounded covariance.

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (Id - B)^{-1}(M_A A + \varepsilon). \tag{C.5}$$

In this context, we use the following notational shorthand,

$$w_\gamma := \left((Id - B)^{-1}_{d_X+1,\cdot} - \gamma^\top (Id - B)^{-1}_{1:d_X,\cdot}\right)^\top, \tag{C.6}$$

such that we can write the residual as a function of both the exogenous noise ε and A as

$$R(\gamma) := Y - \gamma^\top X = w_\gamma^\top(\varepsilon + M_A A), \tag{C.7}$$

under the training distribution. (This identity explains the valley in the loss landscape displayed in Fig. 3: If $d_A \geq 2$, for any parameter γ, there exist an orthogonal intervention direction $\nu \in (w_\gamma^\top M_A)^\perp$, to which the loss is invariant.)

Proposition C.1. *Under Assumption 1,*

$$\begin{aligned} &\ell_{PLS}(X, Y, A; \gamma) \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top w_\gamma, \end{aligned} \tag{C.8}$$

where w_γ is defined by (C.6). If additionally Assumption 2 holds then

$$\begin{aligned} &\ell_{PLS}(X, Y, W; \gamma) \\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] M_A^\top w_\gamma. \end{aligned} \tag{C.9}$$

Proof. The first statement follows from (6) and the observation that

$$\begin{aligned}\mathbb{E}[R(\gamma)A^\top] &= \mathbb{E}[w_\gamma^\top(\varepsilon + M_A A)A^\top] \\ &= w_\gamma^\top \mathbb{E}[\varepsilon A^\top] + w_\gamma^\top M_A \mathbb{E}[AA^\top] \\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top],\end{aligned}$$

where we used $\varepsilon \perp\!\!\!\perp A$. Similarly

$$\begin{aligned}&\ell_{PLS}(X, Y, W; \gamma) \\ &= \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top] \\ &= \mathbb{E}[w_\gamma^\top(\varepsilon + M_A A)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top] \\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]M_A^\top w_\gamma,\end{aligned}$$

where the first equality follows from (6), and the final equality follows from the fact that $\varepsilon \perp\!\!\!\perp W$. □

Proposition C.2. *Under Assumption 1, for any λ and any real, symmetric Ω such that $0 \preceq \mathbb{E}[AA^\top] + \lambda\Omega$, any loss function of the form*

$$\ell(\gamma, \lambda) := \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega M_A^\top w_\gamma, \tag{C.10}$$

where w_γ is defined by (C.6), is equal to the following worst-case loss under bounded perturbations

$$\ell(\gamma, \lambda) = \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2],$$

where

$$C(\lambda) := \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega\}.$$

Proof. We have, making use of the fact that $\varepsilon \perp\!\!\!\perp A$, and $\mathbb{E}[\varepsilon] = 0$

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}\left[(Y - \gamma^\top X)^2\right] \\ &= \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}\left[(w_\gamma^\top(\varepsilon + M_A\nu))^2\right] \\ &= \mathbb{E}\left[(w_\gamma^\top \varepsilon)^2\right] + \sup_{\nu \in C(\lambda)} \mathbb{E}[(w_\gamma^\top M_A \nu)^2] \\ &= \mathbb{E}\left[(w_\gamma^\top \varepsilon)^2\right] + \sup_{\nu \in C(\lambda)} w_\gamma^\top M_A \mathbb{E}[\nu\nu^\top] {M_A}^\top w_\gamma \\ &= \mathbb{E}\left[(w_\gamma^\top \varepsilon)^2\right] + w_\gamma^\top M_A(\mathbb{E}[AA^\top] + \lambda\Omega){M_A}^\top w_\gamma \\ &= \mathbb{E}\left[(w_\gamma^\top \varepsilon)^2\right] + w_\gamma^\top M_A \mathbb{E}[AA^\top] {M_A}^\top w_\gamma\end{aligned}$$

$$
\begin{aligned}
&+ \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \mathbb{E}\left[(w_\gamma^\top(\varepsilon + M_A A))^2\right] + \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \ell(\gamma, \lambda),
\end{aligned}
$$

where in the fifth line we used the definition of $C(\lambda)$. The supremum is achievable even if ν is a deterministic vector, since we can take $\nu \coloneqq \frac{Sb}{\sqrt{b^\top Sb}}$ where $S \coloneqq \mathbb{E}[AA^\top] + \lambda\Omega$ and $b \coloneqq {M_A}^\top w_\gamma$. Then the supremum value is achieved by ν, as $\nu\nu^\top = \frac{Sbb^\top S}{b^\top Sb}$ and $b^\top \nu\nu^\top b = \frac{b^\top Sbb^\top Sb}{b^\top Sb} = b^\top Sb$. To show that $\nu\nu^\top \preceq S$, such that $\nu \in C(\lambda)$, we can take any conformable vector x to see that

$$
\begin{aligned}
x^\top(S - \nu\nu^\top)x &= x^\top Sx - \frac{x^\top Sbb^\top Sx}{b^\top Sb} \\
&= \langle x, x\rangle - \frac{\langle x, b\rangle^2}{\langle b, b\rangle} \\
&\geq 0,
\end{aligned}
$$

where we use the fact that $\langle e, f\rangle \coloneqq e^\top Sf$ defines an inner product, and we apply Cauchy-Schwarz: $\langle x, x\rangle\langle b, b\rangle \geq \langle x, b\rangle^2$. □

In the proofs for Section 3, we will occasionally make use of the following fact, which we prove here to simplify exposition later on.

Proposition C.3. *In the setting of a single proxy (i.e., under Assumptions 1 and 2) let Ω_W be defined as follows*

$$
\Omega_W \coloneqq \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]. \tag{C.11}
$$

Then $\Omega_W \preceq \mathbb{E}[AA^\top]$. Furthermore, if $\mathbb{E}[\varepsilon_W \varepsilon_W^\top]$ is positive definite, then this inequality is strict, that is, $\Omega_W \prec \mathbb{E}[AA^\top]$.

Proof. Recall that $\mathbb{E}[AA^\top]$ and $\mathbb{E}[WW^\top]$ are invertible (and hence positive definite) by assumption. The inequality $\Omega_W \preceq \mathbb{E}[AA^\top]$ is equivalent to showing that $S \coloneqq \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \succeq 0$. Observe that S is the Schur complement of the matrix $K \coloneqq \mathbb{E}\left[\begin{pmatrix} A \\ W \end{pmatrix}\begin{pmatrix} A \\ W \end{pmatrix}^\top\right]$. The matrix K is positive semi-definite (PSD) if and only if $\mathbb{E}[AA^\top]$ is positive definite (true by assumption) and S is PSD (see Zhang [2006, Theorem 1.12b]). Since K is PSD by construction, as the covariance matrix of A, W, this implies that $S \succeq 0$.

Similarly, K is positive definite (PD) if and only if $\mathbb{E}[AA^\top]$ and S are both PD (see Zhang [2006, Theorem 1.12a]). Under the condition that $\mathbb{E}[\varepsilon_W \varepsilon_W^\top]$ is full-rank, then K is PD, and the second inequality follows. □

C.3.2. Proof of additional results

Proof of (9). It follows from Proposition C.1 that

$$\ell_{PLS}(X, Y, A; \gamma) = w_\gamma^\top M_A \Omega_A M_A^\top w_\gamma$$
$$\ell_{PLS}(X, Y, W; \gamma) = w_\gamma^\top M_A \Omega_W M_A^\top w_\gamma,$$

where $\Omega_W := \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]$ and $\Omega_A := \mathbb{E}[AA^\top]$ are both full rank because $\mathbb{E}[AW^\top] = \mathbb{E}[AA^\top]\beta_W$ and by assumptions that $\mathbb{E}[WW^\top], \mathbb{E}[AA^\top]$ and β_W are full rank. Hence both $\ell_{PLS}(X, Y, A; \gamma)$ and $\ell_{PLS}(X, Y, W; \gamma)$ are zero exactly when $w_\gamma^\top M_A = 0$. □

C.3.3. Proof of main results

C.3.3.1. Section 3

Proof of Theorem 1. We use the fact that ε is mean-zero and independent of both A and W. Recall that

$$\ell_{PAR}(W; \gamma, \lambda) = \ell_{LS}(\gamma) + \lambda \ell_{PLS}(W; \gamma),$$

where we suppress the dependence on X, Y in the notation. Letting w_γ be as defined in (C.6), it follows from (C.9) that

$$\ell_{PLS}(X, Y, W; \gamma)$$
$$= w_\gamma^\top M_A \underbrace{\mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]}_{\Omega_W} M_A^\top w_\gamma.$$

The statement then follows from the application of Proposition C.2, and the fact that $\Omega_W \preceq \mathbb{E}[AA^\top]$ (by Proposition C.3), such that $\mathbb{E}[AA^\top] + \lambda\Omega_W \succeq 0$ for all $\lambda \geq -1$. □

Proof of Proposition 1. Recall that the guarantee regions are given by

$$C_A(\lambda) = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\mathbb{E}[AA^\top]\}$$
$$C_W(\lambda) = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega_W\}$$
$$C_{OLS} = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\},$$

where

$$\Omega_W = \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top].$$

The fact that $\mathbb{E}[WW^\top]^{-1} \succ 0$ implies $\Omega_W \succeq 0$, and this implies that $C_{OLS} \subseteq C_W(\lambda)$ for $\lambda \geq 0$. Showing $C_W(\lambda) \subset C_A(\lambda)$ amounts to showing that $\Omega_W \prec \mathbb{E}[AA^\top]$, which holds by Proposition C.3 when $\mathbb{E}[\varepsilon_W\varepsilon_W^\top] \succ 0$.

Next, we prove that C_W is monotonically decreasing in the noise $\mathbb{E}[\varepsilon_W\varepsilon_W^\top]$, in the

sense that if $\mathbb{E}[\varepsilon_W \varepsilon_W^\top] \preceq \mathbb{E}[\eta_W \eta_W^\top]$ then

$$\begin{aligned}&\mathbb{E}_\eta[AW^\top]\mathbb{E}_\eta[WW^\top]^{-1}\mathbb{E}_\eta[WA^\top]\\ &\preceq \mathbb{E}_\varepsilon[AW^\top]\mathbb{E}_\varepsilon[WW^\top]^{-1}\mathbb{E}_\varepsilon[WA^\top],\end{aligned}$$

where $\mathbb{E}_\eta$ is the expectation in the SCM where $W := \beta_W^\top A + \eta_W$ (and similar for $\mathbb{E}_\varepsilon$).

Suppose that $\mathbb{E}[\varepsilon_W \varepsilon_W^\top] \preceq \mathbb{E}[\eta_W \eta_W^\top]$. Then $\mathbb{E}_\eta[WW^\top]^{-1} \preceq \mathbb{E}_\varepsilon[WW^\top]^{-1}$, and since $\mathbb{E}_\eta[AW^\top] = \mathbb{E}_\varepsilon[AW^\top]$, for any vector $x \in \mathbb{R}^{d_A}$ it holds that,

$$\begin{aligned}&(\mathbb{E}_\eta[WA^\top]x)^\top \mathbb{E}_\eta[WW^\top]^{-1}(\mathbb{E}_\eta[WA^\top]x)\\ &\leq (\mathbb{E}_\varepsilon[WA^\top]x)^\top \mathbb{E}_\varepsilon[WW^\top]^{-1}(\mathbb{E}_\varepsilon[WA^\top]x).\end{aligned}$$

This establishes the matrix inequality.

To conclude the proof, suppose that $\mathbb{E}[\varepsilon_W \varepsilon_W^\top] = 0$, $d_A = d_W$ and that β_W has full rank. It then follows that

$$\begin{aligned}\Omega_W &= \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]\\ &= \mathbb{E}[AA^\top]\beta_W(\beta_W^\top \mathbb{E}[AA^\top]\beta_W)^{-1}\beta_W^\top \mathbb{E}[AA^\top]\\ &= \mathbb{E}[AA^\top]\beta_W \beta_W^{-1}\mathbb{E}[AA^\top]^{-1}\beta_W^{\top\,-1}\beta_W^\top \mathbb{E}[AA^\top]\\ &= \mathbb{E}[AA^\top],\end{aligned}$$

such that $C_W(\lambda) = \mathbb{E}[AA^\top] + \lambda\Omega_W = (1+\lambda)\mathbb{E}[AA^\top] = C_A(\lambda)$. □

Proof of Theorem 2. Let w_γ be defined as in (C.6). We can write the population quantity as follows, making use of the fact that ε, ε_Z, and ε_W are jointly independent, and that all errors have zero mean.

$$\begin{aligned}&\ell_\times(W, Z; \gamma)\\ &= \mathbb{E}[(Y - \gamma^\top X)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[Z(Y - \gamma^\top X)^\top]\\ &= \mathbb{E}[w_\gamma^\top(M_A A + \varepsilon)W^\top]\mathbb{E}[ZW^\top]^{-1}\\ &\qquad \cdot \mathbb{E}[Z(A^\top M_A^\top + \varepsilon^\top)w_\gamma]\\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZA^\top]M_A^\top w_\gamma\\ &= w_\gamma^\top M_A \mathbb{E}[A(A^\top\beta_W + \varepsilon_W^\top)]\\ &\qquad \mathbb{E}[(\beta_Z^\top A + \varepsilon_Z)(A^\top\beta_W + \varepsilon_W^\top)]^{-1}\\ &\qquad \mathbb{E}[(\beta_Z^\top A + \varepsilon_Z)A^\top]M_A^\top w_\gamma\\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W\Big(\beta_Z^\top \mathbb{E}[AA^\top]\beta_W\Big)^{-1}\\ &\qquad \beta_Z^\top \mathbb{E}[AA^\top]M_A^\top w_\gamma\\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W \beta_W^{-1}\mathbb{E}[AA^\top]^{-1}(\beta_Z^\top)^{-1}\end{aligned}$$

$$\begin{aligned}
&\beta_Z^\top \mathbb{E}[AA^\top] M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top] \mathbb{E}[AA^\top]^{-1} \mathbb{E}[AA^\top] M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top w_\gamma
\end{aligned}$$

The result follows from Proposition C.1. □

In the main text, we state that the xPAR(W, Z) objective is convex in γ and has a closed form solution. We give the proof here:

Proposition C.4. *Under Assumptions 1, 3 and 4, the loss in (14) is convex in γ, and its minimizer is given by*

$$\begin{aligned}
\gamma^*_{\times PAR} \coloneqq &\Big(2\mathbb{E}[XX^\top] + \lambda(L + L^\top)\Big)^{-1} \\
&\Big(2\mathbb{E}[XY^\top] + \lambda(K_1 + K_2)\Big),
\end{aligned}$$

where we define

$$\begin{aligned}
L &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top], \\
K_1 &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZY^\top] \\
K_2 &:= \mathbb{E}[XZ^\top]\mathbb{E}[WZ^\top]^{-1}\mathbb{E}[WY^\top].
\end{aligned}$$

Proof. By Theorem 2 and (7), $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{AR}(X, Y, A; \gamma, \lambda)$, and the latter is convex in γ, since it is the sum ℓ_{LS}, which is convex, and $\lambda\ell_{PLS}(X, Y, A; \gamma)$, which is a quadratic form by Proposition C.1 and hence convex.

Consequently optimal solution can be found by taking the gradient of $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{LS} + \lambda\ell_\times$ with respect to γ and equating it to 0. Letting $D := \mathbb{E}[ZW^\top]^{-1}$, we can differentiate $\ell_{\times PAR}$ term wise, using (13) to rewrite $\ell_\times$:

$$\begin{aligned}
0 = &\, 2\gamma^\top \mathbb{E}[XX^\top] - 2\mathbb{E}[YX^\top] \\
&- \lambda\mathbb{E}[YW^\top] D \mathbb{E}[ZX^\top] \\
&- \lambda\mathbb{E}[YZ^\top] D^\top \mathbb{E}[WX^\top] \\
&+ \lambda\gamma^\top (L + L^\top),
\end{aligned}$$

where $L := \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top]$. Defining $K_1 := \mathbb{E}[XW^\top] D \mathbb{E}[ZY^\top]$ and $K_2 := \mathbb{E}[XZ^\top] D^\top \mathbb{E}[WY^\top]$, and rearranging, we obtain:

$$\begin{aligned}
&\gamma^\top (2\mathbb{E}[XX^\top] + \lambda(L + L^\top)) \\
&= 2\mathbb{E}[YX^\top] + \lambda(K_1^\top + K_2^\top),
\end{aligned}$$

so by transposing and solving for γ, we get the expression from the statement. □

C.3.3.2. Section 4

Proof of Proposition 2. Let w_γ be defined by (C.6) and for any γ let $b_\gamma^\top := w_\gamma^\top M_A$. We can write the loss as follows

$$
\begin{aligned}
&\mathbb{E}_{do(A:=\nu)}[(Y-\gamma^\top X-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top(\varepsilon+M_A\nu)-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\varepsilon+w_\gamma^\top M_A\nu-\alpha)^2] \\
&\overset{\varepsilon\perp\!\!\!\perp\nu}{=} \mathbb{E}[(w_\gamma^\top\varepsilon)^2]+\mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\varepsilon)^2]+\mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&\qquad -\mathbb{E}[(w_\gamma^\top M_A A)^2]+\mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \ell_{LS}(\gamma)-\mathbb{E}[(b_\gamma^\top A)^2]+\mathbb{E}[(b_\gamma^\top\nu-\alpha)^2] \\
&= \ell_{LS}(\gamma)-b_\gamma^\top\mathbb{E}[AA^\top]b_\gamma^\top \\
&\qquad +b_\gamma^\top\mathbb{E}[\nu\nu^\top]b_\gamma-2\mathbb{E}[b_\gamma^\top\nu]\alpha+\alpha^2 \\
&= \ell_{LS}(\gamma)+b_\gamma^\top\left(\mathbb{E}[\nu\nu^\top]-\mathbb{E}[AA^\top]\right)b_\gamma \\
&\qquad -2\mathbb{E}[b_\gamma^\top\nu]\alpha+\alpha^2 \\
&= \ell_{LS}(\gamma) \\
&\qquad +b_\gamma^\top\left(\mathbb{E}[\nu\nu^\top]-\mathbb{E}[AA^\top]\right)b_\gamma-(b_\gamma^\top\mathbb{E}[\nu])^2 \\
&\qquad +(b_\gamma^\top\mathbb{E}[\nu])^2-2\mathbb{E}[b_\gamma^\top\nu]\alpha+\alpha^2 \\
&= \ell_{LS}(\gamma)+b_\gamma^\top(\Sigma_\nu-\Sigma_A)b_\gamma+\left(b_\gamma^\top\mathbb{E}[\nu]-\alpha\right)^2,
\end{aligned}
$$

where for any value of γ, that minimizing with respect to α yields $\alpha^* = b_\gamma^\top\mathbb{E}[\nu]$, where $b_\gamma^\top = w_\gamma^\top M_A$. Given that we can write the structural relationship $Y-\gamma^\top X = b_\gamma^\top A+w_\gamma^\top\varepsilon$, and knowing that $\mathbb{E}[\varepsilon]=0$ and that $\varepsilon \perp\!\!\!\perp A$, we know that $b_\gamma^\top A$ is the conditional expectation of $R(\gamma)$ given A. □

In the main text, we note that (16) (the objective function ℓ_{TAR}) is convex in γ, α, and has a closed form solution. We prove that result here.

Proposition C.5. *Under Assumption 1, the minimizer $\gamma^*_{TAR}, \alpha^*_{TAR}$ of (16) is given by*

$$
\begin{aligned}
\gamma^* &= \left(\mathbb{E}[XX^\top]+\mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1} \\
&\qquad \left(\mathbb{E}[XY^\top]+\mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right) \\
\alpha^* &= b_{\gamma^*}^\top\mu_\nu,
\end{aligned}
$$

where $\Omega = \mathbb{E}[AA^\top]^{-1}(\Sigma_\nu-\Sigma_A)\mathbb{E}[AA^\top]^{-1}$, and $b_\gamma^\top$ is defined in (15).

Proof of Proposition C.5. Let w_γ be as defined in (C.6) and let $b_\gamma^\top := w_\gamma^\top M_A$. Since $\mathbb{E}[(Y - \gamma^\top X) \mid A] = \mathbb{E}[w_\gamma^\top [M_A A + \varepsilon] \mid A] = b_\gamma^\top A$, for any γ, $b_\gamma^\top$ is the linear regression coefficient of $(Y - \gamma^\top X)$ onto A, so we may write $b_\gamma^\top = \mathbb{E}[(Y - \gamma^\top X)A^\top]\mathbb{E}[AA^\top]^{-1}$. Plugging in the optimal value $\alpha(\gamma) := b_\gamma^\top \mu_\nu$, we obtain

$$\begin{aligned}
&\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha(\gamma)) \\
&= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma \\
&= \ell_{LS}(\gamma) + \mathbb{E}[(Y - \gamma^\top X)A^\top]\Omega\mathbb{E}[A(Y - \gamma^\top X)^\top]
\end{aligned}$$

This objective is convex in γ. The derivative of the loss with respect to γ is

$$-2(\mathbb{E}[(Y - \gamma^\top X)X^\top] + \mathbb{E}[(Y - \gamma^\top X)A^\top]\Omega\mathbb{E}[AX^\top]),$$

and equating to 0 and solving for γ yields

$$\begin{aligned}
\gamma^* = &\left(\mathbb{E}[XX^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1} \\
&\left(\mathbb{E}[XY^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right).
\end{aligned}$$

□

We also claim in the main text that if ν is a constant, then the minimizer of (16) can be found by performing OLS using both X, A as predictors, and then plugging in the known value ν for A in prediction. We prove that result here.

Proof. If ν is a constant, then we can write the first two terms as follows, where w_γ is defined in (C.6).

$$\begin{aligned}
&\ell_{LS} - b_\gamma^\top \Sigma_A b_\gamma \\
&= \mathbb{E}[(w_\gamma^\top (M_A A + \varepsilon))^2] - w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top b_\gamma \\
&= \mathbb{E}[(w_\gamma^\top (M_A A + \varepsilon))^2] - \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&= \mathbb{E}[(w_\gamma^\top \varepsilon)^2]
\end{aligned}$$

which is equivalent to the objective for the loss when Y, X are residualized with respect to A (see Section 8.6 of Rothenhäusler et al. [2021]). By the Frish-Waugh-Lovell theorem [Lovell, 1963, 2008], this yields the same coefficients γ for X as if we had performed regression on X, A together. For this value of γ, $b_\gamma^\top$ is the coefficient that we would obtain for A in the joint regression, because it equals the regression coefficients for $Y - \gamma^\top X$ on A. □

Proof of Proposition 3. We use ν to denote the random shift. Let $\nu \in T(\mu_\nu, \Sigma_\nu)$, or equivalently, let $\nu := \mu_\nu + \delta$, where μ_ν is fixed and δ satisfies the constraint that $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$, where Σ_ν is a symmetric positive definite matrix. Let w_γ be defined by (C.6) and

for any γ let $b_\gamma^\top \coloneqq w_\gamma^\top M_A$. We can write the loss as follows

$$\begin{aligned}
&\sup_{\nu\in T} \mathbb{E}_{do(A:=\nu)}[(Y-\gamma^\top X-\alpha)^2] \\
&= \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top(\varepsilon+M_A\nu)-\alpha)^2] \\
&= \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top\varepsilon+w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\varepsilon)^2] + \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\varepsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&\qquad - \mathbb{E}[(w_\gamma^\top M_A A)^2] + \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \ell_{LS}(\gamma) - \mathbb{E}[(b_\gamma^\top A)^2] + \sup_{\nu\in T} \mathbb{E}[(b_\gamma^\top\nu-\alpha)^2],
\end{aligned}$$

where on the fourth line we used the fact that $\mathbb{E}[\varepsilon\nu]=0$ by the fact that $\nu=\mu_v+\delta$, and δ is independent of ε. In the last line we replaced $w_\gamma^\top M_A$ by $b_\gamma^\top$. We can re-write the last term as follows, where the supremum with respect to δ is constrained in the set $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$

$$\begin{aligned}
&\sup_{\nu\in T} \mathbb{E}[(b_\gamma^\top\nu-\alpha)^2] \\
&= \sup_{\delta:\mathbb{E}[\delta\delta^\top]\preceq\Sigma_\nu} \mathbb{E}[(b_\gamma^\top(\delta+\mu_\nu)-\alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top\delta+b_\gamma^\top\mu_\nu-\alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top\delta)^2] + 2\mathbb{E}[(b_\gamma^\top\delta)](b_\gamma^\top\mu_\nu-\alpha) + \mathbb{E}[(b_\gamma^\top\mu_\nu-\alpha)^2] \\
&= b_\gamma^\top\Sigma_\nu b_\gamma + 2\left\|b_\gamma\right\|_{\Sigma_\nu}\cdot\left|b_\gamma^\top\mu_\nu-\alpha\right| + (b_\gamma^\top\mu_\nu-\alpha)^2,
\end{aligned}$$

where $\left\|b_\gamma\right\|_{\Sigma_\nu} \coloneqq \sqrt{b_\gamma^\top\Sigma_\nu b_\gamma}$ is the norm induced by the inner product defined with respect to Σ_ν. In the last line, we have used the fact that the expression is maximized (subject to the constraint) by the deterministic distribution $\delta_* = \pm\frac{\Sigma_\nu b_\gamma}{\sqrt{b_\gamma^\top\Sigma_\nu b_\gamma}}$ where the sign depends on the sign of $(b_\gamma^\top\mu_\nu-\alpha)$: δ_* satisfies $b_\gamma^\top\delta_*\delta_*^\top b_\gamma = b_\gamma^\top\Sigma_\nu b_\gamma$, maximizing the first term. Further, the second term is also maximized by δ_*, because if any other random or deterministic δ satisfies $|\mathbb{E}b_\gamma^\top\delta| > |b_\gamma^\top\delta_*|$, it follows by Jensens inequality that $\mathbb{E}[(b_\gamma^\top\delta)^2] \geq (\mathbb{E}[(b_\gamma^\top\delta)])^2 > (b_\gamma^\top\delta_*)^2 = b_\gamma^\top\Sigma_\nu b_\gamma$, such that $\mathbb{E}[\delta\delta^\top] \succ \Sigma_\nu$, so δ is not in the set over which the supremum is taken. Consequently, the supremum is attained at δ_*, because δ_* maximizes both terms.

Using this expression for the supremum, we can write the objective as

$$\begin{aligned}
&\sup_{\nu \in T} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\
&= \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma \\
&\qquad + 2 \left\| b_\gamma \right\|_\Sigma \cdot \left| b_\gamma^\top \mu_\nu - \alpha \right| + (b_\gamma^\top \mu_\nu - \alpha)^2,
\end{aligned}$$

for which the optimal choice of α^* is given by $b_\gamma^\top \mu_\nu$, for any γ, and for this choice of α, we can see that $\gamma^* = \arg\min_\gamma \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma$. □

C.4. Targeting with proxies

Definition C.1 (Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ denote the mean of W under intervention, and let $\tilde{\Sigma}_W := \mathrm{cov}_{do(A:=\nu)}(W)$ denote the covariance. We define

$$\begin{aligned}
&\ell_{PTAR}(W; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) \\
&= \ell_{LS}(\gamma) + c_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W \right) c_\gamma + (c_\gamma^\top \tilde{\mu} - \alpha)^2,
\end{aligned} \tag{C.12}$$

where $c_\gamma^\top := \mathbb{E}[R(\gamma) W^\top] {\Sigma_W}^{-1}$.

As mentioned in the main text, (C.12) is not generally equal to (16), and does not generally yield the optimal predictor under the targeted loss. A simple example is given in Proposition C.6.

Proposition C.6. *Assume Assumptions 1, 2, and that $\mathbb{E}[\varepsilon_W \varepsilon_W^\top]$ is full rank. Let $\nu \overset{(d)}{=} A + \eta$ for the deterministic vector $\eta^T = \mathbb{E}[R(\gamma^*_{OLS}) A^\top]$, where $\overset{(d)}{=}$ indicates equality of distribution, and assume $\eta \neq 0$. Then, the minimizers of (16) and (C.12) differ, in that*

$$\alpha^*_{PTAR} < \alpha^*_{TAR}$$

and if $d_W = d_A = 1$, and A has unit variance, then $\frac{\alpha^*_{PTAR}}{\alpha^*_{TAR}} = \rho_W$, where $\rho_W := \beta_W^2 / (\beta_W^2 + \mathbb{E}[\varepsilon_W^2])$.

Proof. The assumption that $\nu = A + \eta$ implies that $\Sigma_\nu - \Sigma_A = 0$, and $\mathbb{E}[\nu] = \eta$. That is, we have changed the mean of the distribution, but not the covariance. This implies

$$\begin{aligned}
\mathbb{E}[\tilde{W}] &= \beta_W^\top \mathbb{E}[\nu] = \beta_W^\top \eta \\
\Sigma_{\tilde{W}} - \Sigma_W &= \beta_W^\top (\Sigma_\nu - \Sigma_A) \beta_W = 0,
\end{aligned}$$

where in the second equation we use the fact that $\Sigma_W = \beta_W^\top \mathbb{E}[A A^\top] \beta_W + \mathbb{E}[\varepsilon_W \varepsilon_W^\top]$ (and similarly for $\Sigma_{\tilde{W}}$), and the ε_W terms cancel in the subtraction. We can then write both

objectives as follows

$$\begin{aligned}
&\ell_{PTAR}(W, \tilde{W}; \gamma, \alpha) \\
&\quad = \ell_{LS}(\gamma) + \left(c_\gamma^\top \beta_W^\top \eta - \alpha\right)^2 \\
&\quad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma)A^T]\beta_W \Sigma_W^{-1} \beta_W^\top \eta - \alpha\right)^2 \\
&\ell_{TAR}(A, \nu; \gamma, \alpha) \\
&\quad = \ell_{LS}(\gamma) + \left(b_\gamma^\top \eta - \alpha\right)^2 \\
&\quad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma)A^T]\Sigma_A^{-1} \eta - \alpha\right)^2
\end{aligned}$$

This gives the optimal value of α in both cases as the value that minimizes the second term

$$\begin{aligned}
\alpha^*_{PTAR} &= \mathbb{E}[R(\gamma^*_{PTAR})A^T](\beta_W \Sigma_W^{-1} \beta_W^\top)\eta \\
\alpha^*_{TAR} &= \mathbb{E}[R(\gamma^*_{TAR})A^T]\Sigma_A^{-1}\eta,
\end{aligned}$$

and since the second term can be made equal to zero by these choices of α, the optimal γ in both cases is identically $\gamma^*_{PTAR} = \gamma^*_{TAR} = \gamma^*_{OLS}$, the value of γ that minimizes the first term $\ell_{LS}(\gamma)$. Hence, we can write the difference between these terms as

$$\begin{aligned}
&\alpha^*_{TAR} - \alpha^*_{PTAR} \\
&= \mathbb{E}[R(\gamma^*_{OLS})A^T](\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top)\mathbb{E}[AR(\gamma^*_{OLS})],
\end{aligned}$$

where we have replaced η with the assumed value of $\mathbb{E}[AR(\gamma^*_{OLS})]$. By assumption, Σ_A is full-rank, so that matrix $\Omega := (\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top)$ is positive definite if and only if $\Sigma_A \Omega \Sigma_A$ is positive definite. Working with this representation, we can see that

$$\begin{aligned}
\Sigma_A \Omega \Sigma_A &= \Sigma_A - \Sigma_A \beta_W \Sigma_W^{-1} \beta_W^T \Sigma_A \\
&= \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \\
&\succ 0,
\end{aligned}$$

where the last line follows from Proposition C.3. In the case where $d_W = d_A = 1$, and A has unit variance, then let $\rho_W = \beta_W^2/(\beta_W^2 + \mathbb{E}[\varepsilon_W^2])$, and observe that

$$\alpha^*_{PTAR} = \eta^2 \rho_W \qquad\qquad \alpha^*_{TAR} = \eta^2.$$

□

Proposition C.6 describes a worst-case mean-shift in A, where η is taken in the direction that maximizes the loss of the OLS solution γ^*_{OLS}. This is also a particularly simple case to analyze for building intuition, because the optimal solution to both (16) and (C.12) is to take $\gamma = \gamma^*_{OLS}$ and to estimate an intercept term α equal to the bias

incurred by the shift in the mean of A. However, the noise in W results in underestimating the impact of the shift, and the gap to the optimal solution depends on the signal-to-variance relationship in W, which (as discussed in Section 3) is not generally identified.

We also prove that the Cross-Proxy Targeted Anchor Regression objective is equal to that of Targeted Anchor Regression.

Theorem C.1. *Under Assumptions 1, 3, and 4, for all* $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,

$$\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2]$$

where $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ *is the mean of* W *under intervention, and* $\tilde{\Sigma}_W$ *is the covariance* $\tilde{\Sigma}_W := \mathrm{cov}_{do(A:=\nu)}(W)$.

Proof of Theorem C.1. We have

$$\begin{aligned} a_\gamma^\top &= \mathbb{E}[R(\gamma)Z^\top](\mathbb{E}[WZ^\top])^{-1} \\ &= \mathbb{E}[R(\gamma)(A^\top\beta_Z + \varepsilon_Z^\top)] \\ &\qquad \mathbb{E}[(\beta_W^\top A + \varepsilon_W)(\beta_Z^\top A + \varepsilon_Z)^\top]^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top]\beta_Z(\beta_W^\top \mathbb{E}[AA^\top]\beta_Z)^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top](\mathbb{E}[AA^\top])^{-1}(\beta_W^\top)^{-1}, \end{aligned}$$

while

$$\begin{aligned} \tilde{\mu} &= \beta_W^\top \mathbb{E}[\nu] \\ \tilde{\Sigma}_W - \Sigma_W &= \beta_W^\top(\Sigma_\nu - \Sigma_A)\beta_W. \end{aligned}$$

With $b_\gamma^\top := w_\gamma^\top M_A$ and w_γ defined by (C.6), we have that

$$\begin{aligned} a_\gamma^\top \tilde{\mu} &= b_\gamma^\top \mathbb{E}[\nu] \\ a_\gamma^\top(\tilde{\Sigma}_W - \Sigma_W)a_\gamma &= b_\gamma^\top(\Sigma_\nu - \Sigma_A)b_\gamma, \end{aligned}$$

which is equivalent to $\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)$ (Definition 4, (16)). The proof is complete by Proposition 2. □

Note that the argument is symmetric for using an observed shift in either Z or W, so it suffices to know the anticipated shift with respect to one proxy.

C.5. Details for experiments

C.5.1. Details of Section 5.1

We outline the details of the simulation experiment in Section 5.1.

Summary We simulate a training data set $\mathcal{D}_{\text{train}}$ from a SCM that induces the structure in Fig. 2, fix $\lambda := 5$ and fit estimators PAR(W) and xPAR(W, Z). We consider the intervention $\mathbb{P}_{do(A:=\nu)}$ with $\nu = (-2.83, 0.35, 0.71)^\top$, and simulate a test data set $\mathcal{D}_{\text{test}}$ from that distribution. We then compute the intervention mean squared prediction error (MSPE) $\hat{\mathbb{E}}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ both for PAR(W) and xPAR(W, Z). We repeat this procedure $m = 10^5$ times for several signal-to-variance ratios x (not including 0), and display the quantiles of the losses in Fig. 5. We also plot the population losses $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ for PAR(W) and xPAR(W, Z), as well as AR(A) and OLS.

Technical details We let $\mathbb{E}[AA^\top] = \beta = \text{Id}$ and $\mathbb{E}[\varepsilon_W \varepsilon_W^\top] = s^2 \,\text{Id}$, such that $W = \beta^\top A + s \cdot \varepsilon_W$. Then Ω_W as defined in (11) simplifies to

$$\begin{aligned}\Omega_W &= \mathbb{E}[AA^\top]\beta(\beta^\top \mathbb{E}[AA^\top]\beta + \mathbb{E}[\varepsilon_W \varepsilon_W^\top])^{-1}\beta^\top \mathbb{E}[AA^\top] \\ &= \frac{1}{1+s^2}\,\text{Id}\,.\end{aligned}$$

We call $x = (1+s^2)^{-1}$ the signal-to-variance ratio, and we can obtain a given signal-to-variance ratio x, by setting $s = \sqrt{(1-x)/x}$.

For each $n \in \{150, 500\}$ and signal-to-variance ratio $x \in \{1/20, 2/20, \ldots, 20/20\}$, we set $s = \sqrt{(1-x)/x}$ and sample a data set $\mathcal{D}^i_{n,s}$ for $i = 1, \ldots, 5000$, each with sample size n, from the structural equations:

$$\begin{aligned} A &:= \varepsilon_A \\ W &:= A + s \cdot \varepsilon_W \\ Z &:= A + s \cdot \varepsilon_Z \\ (Y, X, H) &:= (\text{Id} - B)^{-1}(MA + \varepsilon), \end{aligned} \tag{C.13}$$

where $d_A = d_W = d_Z = d_X = 3$, $d_Y = d_H = 1$. M and B are given by

$$M = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 2 & 1 \\ -1 & 3 & 0 \\ 2 & 2 & -3 \\ 0 & -2 & 2 \end{pmatrix}, B = \begin{pmatrix} 0 & -2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and all noise variables are i.i.d., $\varepsilon_A, \varepsilon_W, \varepsilon_Z, \varepsilon \sim \mathcal{N}(0, \text{Id})$. For every combination (n, s) we have 5000 data sets $\mathcal{D}^i_{n,s}$, $i = 1, \ldots, 5000$. For each data set, we compute the proxy estimators $\gamma^i_{n,s,W}$ and $\gamma^i_{n,s,W;Z}$, using one or two proxies respectively, and we simulate 5000 corresponding test data sets of size n from $\mathbb{P}_{do(A:=\nu)}$ (using the structural equations above, except for changing the assignment for A to $A := \nu$). The prediction MSE for the i'th test data set is then $\frac{1}{n}\sum_{j=1}^n (Y_j - \gamma^\top X_j)^2$, resulting in 5000 values of the MSE for each combination of (n, s).

At each combination of (n, s) we plot the median by a line of the estimated worst case losses, and by a shaded region indicate the interval between the 25% and 75% quantiles

of the observed distribution. We plot the median instead of the mean since for small x, $s^2 = \frac{1-x}{x}$ is large, and especially for $\mathrm{WCL}^i_{n,s}(W, Z)$ and $n = 150$, the mean will be driven very much by outliers for small x.

The population versions of losses for any s is computed first by computing the population estimators γ from the parameter matrices M, B, and then computing the loss at ν by $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2] = w_\gamma^\top M \nu\nu^\top M^\top w_\gamma + w_\gamma^\top \mathbb{E}[\varepsilon\varepsilon^\top] w_\gamma$.

C.5.2. Details of Section 5.2

We outline the details of the simulation experiment in Section 5.2.

Summary We analyze the effect of applying anchor regression with one proxy, PAR(W), when the signal-to-variance ratio is potentially misspecified. To do so, we simulate data from the same SCM as in Section 5.1 ($n = 10^4$), and in particular from a range of true (unknown) signal-to-variance ratios $x \in (0, 1]$. To each data set, we apply anchor regression with one proxy, PAR(W), and with $\lambda := 5$. We further assume the signal-to-variance ratio to be 40% – independently of its true value. This means, by Theorem 1, that we assume that PAR(W) minimizes the worst case mean squared prediction error (MSPE) over the region $C := \{\nu\nu^\top \preceq (1 + 0.4 \cdot \lambda)\mathbb{E}[AA^\top]\}$, with the worst case MSPE for being equal to the optimal value of the PAR(W) objective function. If $x = 0.4$, then PAR(W) indeed minimizes the worst case MSPE over C and the estimated worst case MSPE over C is close to the actual worst case MSPE over C. But if $x \neq 0.4$, the estimator minimizes the worst case MSPE over a different set, and then expect that the true worst case MSPE over C differs from its estimate. Fig. 6 shows that this is indeed the case: We observe that if the true signal-to-variance ratio is larger than the assumed 40%, our estimate of the MSPE is too conservative. On the contrary, if the true signal-to-variance ratio is smaller than assumed, our estimates of the MSPE over C are too small, meaning that we underestimate the worst case MSPE in the region C.

Technical details For a fixed signal-to-variance ratio x, we simulate a training data set $\mathcal{D}_{\text{train}}$ ($n = 10^4$) from the same procedure as in Appendix C.5.1, i.e. using the structural equations in (C.13), and with the same parameters M and B. We fit the PAR(W) estimator to the data using $\lambda := 5$, and the estimated worst case mean squared prediction error (MSPE) over C is then the value of the objective function in the estimated parameter (by Theorem 1).

To find the actual worst case MSPE over C for a given estimator λ, we use the fact from (C.7) that

$$\mathbb{E}_{do(A:=v)}[(R - \gamma^\top X)^2] = (b_\gamma^\top v)^2 + w_\gamma^\top w_\gamma, \tag{C.14}$$

where we use that $\mathbb{E}[\varepsilon\varepsilon^\top] = \mathrm{Id}$, w_γ is given by (C.6) and $b_\gamma^\top = w_\gamma^\top M_A$. The second term doesn't depend on v, and since C is spherical, the worst case MSPE over C is attained in the direction $v \propto b_\gamma$, with v normalized such that $\|v\|^2 = (1 + 0.4 \cdot \lambda)$ (that is v lies

on the boundary of C). Using the known M and B, we compute w_γ, b_γ, and the actual worst case MSPE over C is given by (C.14) plugging in $v = b_\gamma \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_\gamma\|$.

We compute also the worst case MSPE over C when using an OLS estimator for the prediction. We fit $\hat{\gamma}_{OLS}$ from $\mathcal{D}_{train}$, and, as for the actual MSPE of PAR(W), the worst case MSPE over C using OLS can be computed, by computing vectors $b_{\hat{\gamma}_{OLS}}, w_{\hat{\gamma}_{OLS}}$. Again the worst case MSPE over C using $\hat{\gamma}_{OLS}$ is attained by setting $v = b_{\hat{\gamma}_{OLS}} \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_{\hat{\gamma}_{OLS}}\|$ and plugging v, $b_{\hat{\gamma}_{OLS}}$ and $w_{\hat{\gamma}_{OLS}}$ into (C.14).

For every signal-to-variance ratio $x \in \{1/20, \dots, 20/20\}$, we repeat the procedure $m = 1000$ times, for each computing the estimated and actual MSPEs. In Fig. 6 we plot the median MSPE as well as the interval from the 25% quantile to the 75% quantile.

C.5.3. Details of Section 5.3

We outline the details of the simulation experiment in Section 5.3.

Summary We demonstrate the ability of Proxy Anchor Regression to select invariant predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data sets ($n = 10^5$) from a SCM with the structure shown in Fig. 7 (top), where one anchor, A_1, is a parent of the causal predictors, while the other A_2 is a parent of the anti-causal predictors.

We consider two identically distributed noisy proxies W, Z of $A := (A_1, A_2)$. The challenge, in this scenario, is that A_2 is measured with significantly more noise than A_1, across both proxies. As a consequence, proxy anchor regression with one proxy, PAR(W), puts more weight on anti-causal features: the noise in W is mistaken for fluctuations in A_2, resulting in $X_{\text{anti-causal}}$ mistakenly appearing invariant to shifts in A_2. In contrast, when two proxies W, Z are available, the estimator xPAR(W, Z) asymptotically equals that of anchor regression with observed anchors, and its regression coefficients puts more weight on the causal predictors; see Fig. 7 (bottom).

Technical details With $d_{A_1} = d_{A_2} = d_W = d_W = 6$, $d_{X_{\text{causal}}} = d_{X_{\text{anti-causal}}} = 3$ and $d_Y = 1$, we simulate data from the SCM in Fig. 7 (top) which amounts to simulating from the following structural equations:

$$\begin{aligned}
A_1 &:= \varepsilon_{A_1} \\
A_2 &:= \varepsilon_{A_2} \\
W &:= (A_1, A_2)^\top + (\varepsilon_{W,1}, \varepsilon_{W,2})^\top \\
Z &:= (A_1, A_2)^\top + (\varepsilon_{Z,1}, \varepsilon_{Z,2})^\top \\
X_{\text{causal}} &:= M_1 A_1 + \varepsilon_{X_{\text{causal}}} \\
Y &:= \gamma_{\text{causal}}^\top X_{\text{causal}} + \varepsilon_Y \\
X_2 &:= M_2 A_2 + \gamma_{\text{anti-causal}} Y + \varepsilon_{X_{\text{anti-causal}}}.
\end{aligned}$$

Here $M_1 \in \mathbb{R}^{d_{X_{\text{causal}}} \times d_{A_1}}$ and $M_2 \in \mathbb{R}^{d_{X_{\text{anti-causal}}} \times d_{A_2}}$ are matrices with 1 in every entry, $\gamma_{\text{causal}} = (1/4, 1/4, 1/4)^\top$ and $\gamma_{\text{anti-causal}} = (4, 4, 4)^\top$ (such that the regression coefficients of Y onto $X_{\text{causal}}, X_{\text{anti-causal}}$ are of similar magnitudes). All noise terms are independent and $\varepsilon_{A_1}, \varepsilon_{A_2}, \varepsilon_{X_{\text{causal}}}, \varepsilon_{X_{\text{anti-causal}}}, \varepsilon_Y \sim \mathcal{N}(0, \text{Id})$, and $\varepsilon_{W,1}, \varepsilon_{Z,1} \sim \mathcal{N}(0, \text{Id})$, $\varepsilon_{W,2}, \varepsilon_{Z,2} \sim \mathcal{N}(0, 3^2 \cdot \text{Id})$.

We simulate a data set $\mathcal{D}$ ($n = 10^5$) from these structural equations, and fit the proxy anchor regression estimators $\gamma(W)$ and $\gamma(W, Z)$ from Section 3. We repeat this $m = 10^4$ times, and display the mean absolute value of the regression coefficients (that is the entries of the vectors $\gamma(W)$ and $\gamma(W, Z)$) in Fig. 7 (bottom), as well as the standard deviation of the absolute value of the regression coefficients as error bars.

C.5.4. Details of Section 5.4

Summary We demonstrate the trade-off made by Targeted Anchor Regression (TAR) versus Anchor Regression (AR), considering the case when A is observed for simplicity. We simulate training data and fit estimators γ_{OLS}, γ_{AR} and γ_{TAR}, where γ_{TAR} is targeted to a particular mean and covariance of a random intervention $do(A := \nu)$, and we select λ for γ_{AR} such that this intervention is contained within $C_A(\lambda)$. We then simulate test data from two distributions: $\mathbb{P}_{do(A:=\nu)}$ (i.e., the shift occurs), and $\mathbb{P}$ (where it does not), and evaluate the mean squared prediction error (MSPE). The results are shown in Fig. 8, and demonstrated that TAR performs better than AR and OLS in the first scenario, but this comes at the cost of worse performance on the training distribution.

Technical details The entire procedure below produces a prediction MSE for each of three methods and two settings, and we repeat this $m = 10^5$ times, to produce the histograms of MSEs shown in Fig. 8.

We simulate a training data set $\mathcal{D}_{\text{train}}$ ($n_{\text{train}} = 10^5$) from the structural equations

$$\begin{aligned} A &:= \varepsilon_A \\ (Y, X, H) &:= (\text{Id} - B)^{-1}(MA + \varepsilon), \end{aligned}$$

where $d_A = d_X = 2$ and $d_Y = d_H = 1$, $\varepsilon_A, \varepsilon \sim \mathcal{N}(0, \text{Id})$ and M and B were selected by a simulation resulting in:

$$M = \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B = \begin{pmatrix} 0 & -0.06 & 0.07 & 0.04 \\ 0.05 & 0 & 0.19 & 0.03 \\ 0.11 & -0.11 & 0 & 0.1 \\ -0.02 & 0.02 & 0.09 & 0 \end{pmatrix}.$$

We consider the target distribution $do(A := \kappa^\top A + \eta)$ where

$$\kappa = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{pmatrix}, \eta = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and so we fit the targeted AR estimator $(\gamma_{\text{targeted-AR}}, \alpha_{\text{targeted-AR}})$ from (16), where the

covariance of the anticipated shift is given by $\Sigma_\nu := \kappa^\top \mathbb{E}[AA^\top]\kappa$, and the mean shift is simply η. We also fit OLS estimates $\gamma_{\text{OLS}}(X, Y)$ and $\gamma_{\text{AR}}(X, Y, A)$ where for AR we select λ such that $(1 + \lambda)$ equals the largest eigenvalue of $\kappa^\top \mathbb{E}[AA^\top]\kappa + \eta\eta^\top$, such that $\mathbb{E}[(\kappa^\top A + \eta)(\kappa^\top A + \eta)^\top] \preceq (1 + \lambda)\mathbb{E}[AA^\top]$.

We then simulate a test data set ($n_{\text{test}} = 10^5$) both from 1) the training distribution (i.e. same simulation procedure as for the training set) or 2) by changing the structural equation for A to $A := \kappa^\top \varepsilon_A + \eta$, and keeping all other quantities as for the simulation of training data (i.e. the test distribution is the anticipated distribution). We evaluate the prediction MSE on each of the data sets by $\frac{1}{n_{\text{test}}} \sum_j (Y_j - \gamma^\top X_j)^2$ (including the term $\alpha_{\text{targeted-AR}}$ for the targeted AR).

C.5.5. Details of Section 6

Features The dataset contains time-stamps as well as season indicators, which we do not use anywhere as features. The remaining features are Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), Combined wind direction (NE, NW, SE, SW, or CV, indicating calm and variable), Cumulated wind speed (m/s), Hourly precipitation (mm), and Cumulated precipitation (mm).

Data Processing Each city has PM2.5 readings from multiple sites, which we average to get a single reading, and we take a log transformation. For Precipitation (Cumulative) we subtract off the (current hour) precipitation to avoid co-linearity. We take a log transformation of the variable for Wind Speed, Precipitation (Hourly) and Precipitation (Cumulative), due to skewness. We drop all rows that contain any missing data.

Proxies (Temperature) We use temperature as our proxy variable, and treat it as unavailable at test time. We construct two synthetic proxies of temperature to serve as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\text{var}(A)/\text{var}(W) = 0.9$. This results in different standard deviations of the Gaussian noise across different environments, because of differences in the training distributions across training seasons and cities. The standard error of the noise varies between 2 and 5 degrees, to maintain the same signal-to-variance ratio.

Training Details (PAR, xPAR) For the distributional robustness approaches described in Section 3, we choose $\lambda \in [0, 40]$ by leave-one-group-out cross-validation on the three training seasons, using the first year (2013) of data. For Proxy Anchor Regression using Temperature directly, there is heterogeneity in the cross-validated choice of λ: In 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent to OLS. We saw a similar result when the maximum value of λ was 20, and increased the maximum limit to 40 without seeing much difference, so we did not increase it further. Concretely, with λ in [0, 20], there are some scenarios where PAR (TempC) has slightly worse or slightly better MSE (vs. λ in [0, 40]), but the differences are all less than 0.001. The only observable difference in Table 1 when running with λ in [0, 20] is that the "best" performance is -0.040 ($\lambda = 20$), as opposed to -0.041 ($\lambda = 40$)

Table C.1.: MSE (lower is better) over 20 scenarios consisting of five cities and four held-out seasons. Average difference to OLS estimator (lower is better) given in the second column, and minimum / maximum difference in remaining columns.

Estimator	Mean	Diff	Min	Max
OLS	0.457			
OLS (TempC)	0.455	-0.002	-0.028	0.026
OLS + Est. Bias	0.474	0.018	-0.072	0.150
PAR (TempC)	0.454	-0.003	-0.041	0.006
PAR (W)	0.454	-0.002	-0.037	0.006
xPAR (W, Z)	0.454	-0.003	-0.039	0.007
PTAR	0.450	-0.007	-0.061	0.002
PTAR (W)	0.452	-0.005	-0.038	0.001
xPTAR (W, Z)	0.450	-0.007	-0.059	0.003

[where lower is better, rounded to nearest 0.001]. For Proxy Anchor Regression using W and for Cross-Proxy Anchor Regression (xPAR) using W, Z together, we use the same values of λ as above, for comparability.

Training Details (PTAR, xPTAR) For the targeted approaches described in Section 4, we use the mean and variance of the temperature in the test distribution to target our predictors, and similarly use the distribution of the proxies when using Proxy Targeted Anchor Regression (PTAR) with W and Cross-Proxy TAR (xPTAR) with W, Z. Note that xPTAR (unlike xPAR) is asymmetric in the proxies, but in this case the proxies are distributed identically.

Benchmarks As described in the main text, our primary benchmark is OLS, trained on the three training seasons, evaluated on the held-out season. We also include two other baselines: First, OLS that has access to temperature during both train and test, which we denote OLS (TempC), and OLS that includes temperature during training, and attempts to estimate a bias term by plugging in the mean (test) value for temperature during prediction.

In Table C.1 we give the full results over all 20 scenarios, which includes the 11 scenarios where $\lambda = 0$ is chosen by cross-validation, rendering the PAR and xPAR solutions equivalent to OLS.

Regularization paths In Fig. C.4 we have shown how the solution in the "best" scenario differs for Proxy Anchor Regression (PAR) with $\lambda = 40$ versus OLS (i.e., $\lambda = 0$). In Fig. C.5, we show how the coefficients change in-between these two extremes: for every integer value of λ in [0, 40] we show the difference in the PAR vs. OLS coefficients

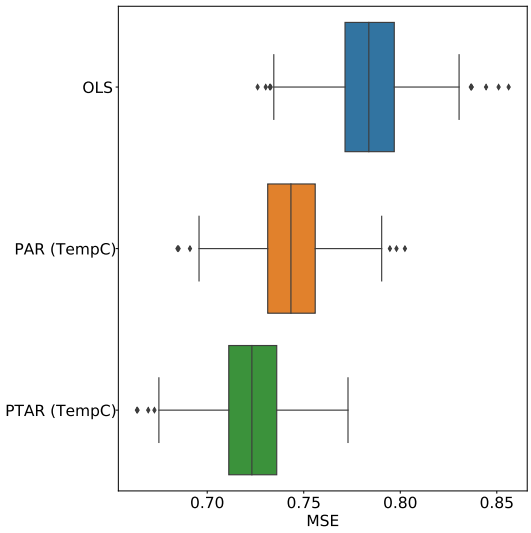


Figure C.3.: Best performance for Proxy Anchor Regression (PAR) and Proxy Targeted AR (PTAR), corresponding to Summer in Beijing. Variance estimates generated by bootstrapping the test residuals of the fitted models.

for each feature. Increasing λ further does not make a significant difference for this particular example.

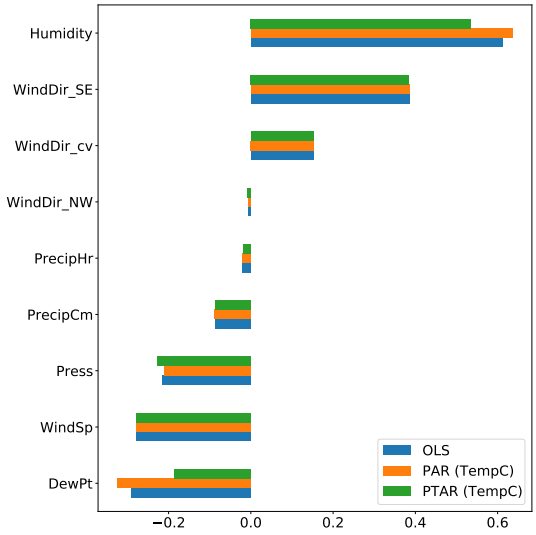


Figure C.4.: Comparison of learned coefficients. All variables were standardized to unit variance. The intercept for OLS and AR is the same (by construction) at $\alpha = 4.087$ while the intercept for TAR is lower at $\alpha = 3.885$.

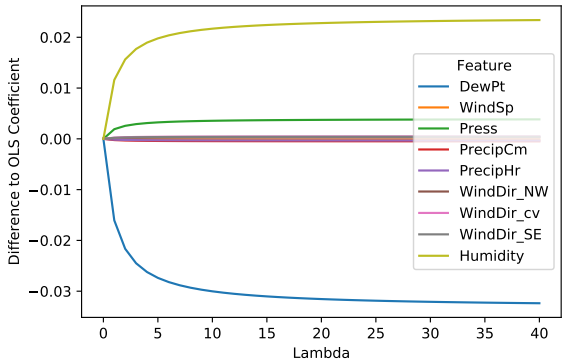


Figure C.5.: Coefficient path, showing the difference between the PAR and OLS coefficients in Fig. C.4 for different values of λ.

C.6. Additional experiment: Signal-to-variance ratio

To examine the effect of the signal strengths β_W and β_Z, we scale the signals $\beta_{W,s} = \beta_{Z,s} = s\,\mathrm{Id}$ for $s \in \{0, \sqrt{2/3}, 0.8\}$, which for the single proxy estimator $\hat{\gamma}_{\mathrm{PAR}}$ amounts to optimizing over worst case loss in the robustness regions $C(\lambda) = \{vv\top \preceq (1+\lambda\frac{s^2}{1+s^2})\,\mathrm{Id}\}$.

For $s \in \{1, 3\}$, such that the signal-to-variance ratio $\frac{s^2}{1+s^2}$ equals either 10% or 50%, we simulate a training data set $\mathcal{D}_{\mathrm{train}}$ with two proxies W and Z from the structural equations $A := \varepsilon_A, (X^\top, Y^\top, H^\top)^\top := (1-B)^{-1}(M_a A + \varepsilon), W := \beta_{W,s}^\top A + \varepsilon_W$ and $Z := \beta_{Z,s}^\top A + \varepsilon_Z$ where all noise terms are i.i.d with unit covariance and M_A, B are given by:

$$M := \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B := \begin{pmatrix} 0 & -0.57 & 0.73 & 0.37 \\ 0.53 & 0 & 1.91 & 0.33 \\ 1.14 & -1.13 & 0 & 0.96 \\ -0.22 & 0.16 & 0.87 & 0 \end{pmatrix}.$$

Since for this experiment we are not interested in finite sample properties of the estimators, we use sample size $n = 10^7$.

For each data set we fit estimators $\hat{\gamma}_{\mathrm{PAR}(W)}$ (using only one proxy), $\hat{\gamma}_{\mathrm{xPAR(W,\,Z)}}$ (using both proxies), $\hat{\gamma}_{\mathrm{AR}(A)}$, and $\hat{\gamma}_{\mathrm{OLS}}$, and evaluate the estimators at data sampled from interventional distributions $\mathbb{P}_{do(A:=v)}$ for several interventions v of increasing strength (i.e. increasing distance from $\mathbb{E}[A] = 0$).

As the signal to variance ratio increases, the PAR(W) loss approaches the AR(A). Further we observe that xPAR(W, Z) coincides with the $AR(A)$ estimator for both signal-to-variance levels. This is illustrated in Fig. C.6.

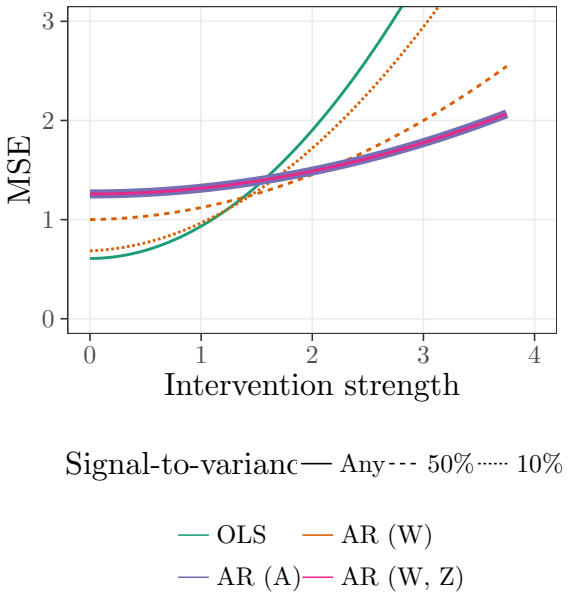


Figure C.6.: Anchor and proxy estimators for different levels of signal-to-variance ratio $\beta(\mathbb{E}[WW^\top])^{-1}\beta^\top$. A training data set ($n = 10^7$) with two proxies W, Z is simulated and the estimators $\hat{\gamma}_{\text{PAR}(A)}$, $\hat{\gamma}_{\text{xPAR}(W,Z)}$, $\hat{\gamma}_{\text{AR}(A)}$, and $\hat{\gamma}_{\text{OLS}}$ are fitted using a fixed λ. Interventions v of increasing strength is sampled, and for each a new data set ($n = 10^5$) is sampled from $\mathbb{P}^{do(A:=v)}$, and for each estimator $\hat{\gamma}$, the prediction mean squared error $\mathbb{E}_{do(A:=(v_1,v_2))}[(Y - \hat{\gamma}^\top X)^2]$ is computed. This procedure is repeated for signal-to-variance ratios 10% and 50%.

D. Appendix to Invariant Policy Learning: A Causal Perspective

D.1. Pearl's d-separation

Definition D.1 (Pearl's d-separation Pearl [2009], Peters et al. [2017])**.** Let $\mathcal{G}$ be directed acyclic graph (DAG) with nodes $\mathbf{V}$. Let $V_i, V_m \in \mathbf{V}$ and $\mathbf{S} \subseteq \mathbf{V} \setminus \{V_i, V_m\}$. A path between nodes V_i and V_m is said to be *blocked by a set* $\mathbf{S}$ if there exists a node $V_k \in \mathbf{V}$ such that one of the following holds:

1. $V_k \in \mathbf{S}$ and

$$\begin{aligned} V_{k-1} \to V_k \to V_{k+1} \\ \text{or } V_{k-1} \leftarrow V_k \leftarrow V_{k+1} \\ \text{or } V_{k-1} \leftarrow V_k \to V_{k+1}, \end{aligned}$$

2. neither V_k nor any of its descendants is in $\mathbf{S}$ and

$$V_{k-1} \to V_k \leftarrow V_{k+1}.$$

For any three disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{S} \subseteq \mathbf{V}$ of nodes in $\mathcal{G}$, we say that $\mathbf{A}$ and $\mathbf{B}$ are *d-separated* by $\mathbf{S}$, denoted by $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$.

(This formulation is taken from Peters et al. [2017].)

D.2. Consistency in the Unconfounded Setting

Let $\widehat{Q}_n$ be an estimator of the conditional mean $\mathbb{E}^{\pi_a}[R \mid X]$ that is based on n independent observations (X_i, A_i, R_i) from potentially different environments. The following proposition shows that such an approach indeed yields a consistent estimate of an optimal policy given that $\widehat{Q}_n$ is consistent.

Proposition D.1. *Assume Setting 2 and let $\widehat{Q}_n$ be a uniformly consistent estimator of $Q^{\mathcal{E}^{\text{obs}}}$, that is, for all $a \in \mathcal{A}$ it holds that*

$$\lim_{n \to \infty} \mathbb{E}_D \left[\sup_{x \in \boldsymbol{\mathcal{X}}} \left| \widehat{Q}_n(x, a) - Q^{\mathcal{E}^{\text{obs}}}(x, a) \right| \right] = 0,$$

where $\mathbb{E}_D$ is an expectation over the n observations (X_i, A_i, R_i) used to estimate $\widehat{Q}_n$. Let $\widehat{\pi}_n$ be any policy that maximizes $\widehat{Q}_n$, i.e., for all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$ it holds that

$$\widehat{\pi}_n(a|x) > 0 \implies a \in \arg\max_{a' \in \mathcal{A}} \widehat{Q}_n(x, a').$$

Then, the robust policy value converges towards its optimal value, that is

$$\lim_{n\to\infty} \mathbb{E}_D \left[\left| V^{\mathcal{E}}(\widehat{\pi}_n) - \max_{\pi \in \Pi} V^{\mathcal{E}}(\pi) \right| \right] = 0.$$

Proof. See Appendix D.3.2. □

The same argument would work if instead of pooling, one considers only a single environment. In practice, however, one would make use of all available data. Whether it is possible to construct a uniformly consistent estimator $\widehat{Q}_n$ depends on the model class that can be assumed in the structural assignment of R, and on the policy used in generating the observations. For example, in the case of additive confounding and noise such as $f(X, U, A, \varepsilon_R) = f_1(X, A) + f_2(U, \varepsilon_R)$ with f_1 and f_2 in some function classes and a policy π that has full support, (i.e., $\forall a \in \mathcal{A}, x \in \mathcal{X} : \pi(a \mid x) > 0$), one can consider a least squares estimator of the form

$$\widehat{Q}_n^{\mathcal{E}^{\text{obs}}} := \arg\min_{f_1} \frac{1}{n} \sum_{i=1}^{n} (f_1(X_i, A_i) - R_i)^2.$$

The assumptions of Proposition D.1 are then satisfied under further constraints on the function class and noise distributions, e.g., linear functions, Gaussian noise, and bounded domains.

D.3. Proofs

D.3.1. Proof of Theorem 1

Proof. We begin by showing that the model class in Setting 2 satisfies an invariance property. Let $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $x \in \boldsymbol{\mathcal{X}}$ be arbitrary. By using the explicit SCM structure from Setting 2, it holds that

$$\begin{aligned} &\mathbb{E}^{\pi_a, e}\left[R \mid X = x\right] \\ &\quad = \mathbb{E}^{\pi_a, e}\left[f(X, s(X, \varepsilon_U), A, \varepsilon_R) \mid X = x\right]. \end{aligned}$$

Since we assume there is no hidden confounding, it holds that $\varepsilon_U \perp\!\!\!\perp X$ which implies that

$$\begin{aligned} &\mathbb{E}^{\pi_a, e}\left[f(X, s(X, \varepsilon_U), A, \varepsilon_R) \mid X = x\right] \\ &\quad = \mathbb{E}_{\varepsilon_U, \varepsilon_R}\left[f(x, s(x, \varepsilon_U), a, \varepsilon_R)\right] \end{aligned}$$

and hence $\mathbb{E}^{\pi_a,e}\left[R \mid X = x\right]$ does not depend on the environment. This, in particular, implies that for all $e \in \mathcal{E}$, all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$, it holds that

$$\begin{aligned} Q^{\mathcal{E}^{\text{obs}}}(x,a) &= \tfrac{1}{|\mathcal{E}^{\text{obs}}|} \sum_{f \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi_a,f}[R \mid X = x] \\ &= \mathbb{E}^{\pi_a,e}[R \mid X = x]. \end{aligned} \tag{D.1}$$

We thus have for all policies $\pi \in \Pi$ and for all $x \in \boldsymbol{\mathcal{X}}$ that

$$\begin{aligned} &\max_{a \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a) \\ &\qquad = \max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X = x] \\ &\qquad \geq \sum_{a \in \mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X = x]\pi(a \mid x) \\ &\qquad = \sum_{a \in \mathcal{A}} \mathbb{E}^{\pi,e}[R \mid X = x, A = a]\pi(a \mid x) \\ &\qquad = \mathbb{E}^{\pi,e}[R \mid X = x]. \end{aligned} \tag{D.2}$$

Next, take the expectation over X on both sides to get

$$\begin{aligned} \mathbb{E}^e\left[\max_{a \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(X,a)\right] &\geq \mathbb{E}^e\left[\mathbb{E}^{\pi,e}[R \mid X]\right] \\ &= \mathbb{E}^{\pi,e}\left[R\right]. \end{aligned}$$

Finally, taking the infimum over $e \in \mathcal{E}$ leads to

$$\inf_{e \in \mathcal{E}} \mathbb{E}^e\left[\max_{a \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(X,a)\right] \geq \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi,e}\left[R\right]. \tag{D.3}$$

Let π^* be a policy such that for all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$

$$\pi^*(a|x) > 0 \implies a \in \arg\max_{a' \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a'). \tag{D.4}$$

Then π^* satisfies, for all $e \in \mathcal{E}$,

$$\mathbb{E}^{\pi^*,e}\left[R\right] = \mathbb{E}^e\left[\max_{a \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(X,a)\right].$$

Therefore (D.3) implies

$$\pi^* \in \arg\max_{\pi \in \Pi} \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi,e}\left[R\right],$$

which completes the proof of Theorem 1. □

D.3.2. Proof of Proposition D.1

Proof. Define for all $n \in \mathbb{N}$ the term

$$c(n) \coloneqq \max_{a \in \mathcal{A}} \sup_{x \in \mathcal{X}} |Q^{\mathcal{E}^{\text{obs}}}(x, a) - \widehat{Q}_n(x, a)|.$$

As $\mathcal{A}$ is assumed to be finite and because $\widehat{Q}_n$ is assumed to be uniformly consistent, it holds that

$$\lim_{n \to \infty} \mathbb{E}_D[c(n)] = 0. \tag{D.5}$$

Moreover, as shown in (D.1), in the proof of Theorem 1, we know that for all $e \in \mathcal{E}$, all $a \in \mathcal{A}$ and all $x \in \mathcal{X}$ it holds that

$$Q^{\mathcal{E}^{\text{obs}}}(x, a) = \mathbb{E}^{\pi_a, e}[R \mid X = x].$$

This implies that for all $x \in \boldsymbol{\mathcal{X}}$ and all $e \in \mathcal{E}$ it holds that

$$\begin{aligned}
&\mathbb{E}^{\widehat{\pi}_n, e}[R \mid X = x] \\
&\quad = \sum_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X = x] \widehat{\pi}_n(a|x) \\
&\quad = \sum_{a \in \mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x, a) \widehat{\pi}_n(a|x) \\
&\quad = \sum_{a \in \mathcal{A}} \widehat{Q}_n(x, a) \widehat{\pi}_n(a|x) \\
&\qquad + \sum_{a \in \mathcal{A}} (Q^{\mathcal{E}^{\text{obs}}}(x, a) - \widehat{Q}_n(x, a)) \widehat{\pi}_n(a|x).
\end{aligned} \tag{D.6}$$

Each of the sums only contains one terms, since $\widehat{\pi}_n$ puts all mass on a single action. Next, observe that

$$\begin{aligned}
&\left| \sum_{a \in \mathcal{A}} (Q^{\mathcal{E}^{\text{obs}}}(x, a) - \widehat{Q}_n(x, a)) \widehat{\pi}_n(a|x) \right| \\
&\quad \leq \sum_{a \in \mathcal{A}} \left| Q^{\mathcal{E}^{\text{obs}}}(x, a) - \widehat{Q}_n(x, a) \right| \widehat{\pi}_n(a|x) \\
&\quad \leq c(n)
\end{aligned} \tag{D.7}$$

and

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} \widehat{Q}_n(x,a)\widehat{\pi}_n(a|x) \\
&\quad = \max_{a\in\mathcal{A}} \widehat{Q}_n(x,a) \\
&\quad = \max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a) \\
&\qquad\qquad + (\max_{a\in\mathcal{A}} \widehat{Q}_n(x,a) - \max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a)).
\end{aligned} \tag{D.8}$$

Using (D.6) to (D.8) together with the triangle inequality yields

$$\begin{aligned}
&\left|\mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x] - \max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a)\right| \\
&\quad = \Big| \max_{a\in\mathcal{A}} \widehat{Q}_n(x,a) - \max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a) \\
&\qquad\qquad + \sum_{a\in\mathcal{A}} (Q^{\mathcal{E}^{\text{obs}}}(x,a) - \widehat{Q}_n(x,a))\widehat{\pi}_n(a|x)\Big| \\
&\quad \leq 2c(n).
\end{aligned}$$

This in particular implies that for all $e \in \mathcal{E}$ and all $x \in \boldsymbol{\mathcal{X}}$ it holds that

$$\max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a) - 2c(n) \leq \mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x]$$

and that

$$\mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x] \leq \max_{a\in\mathcal{A}} Q^{\mathcal{E}^{\text{obs}}}(x,a) + 2c(n).$$

Taking the expectation over X and the infimum over $\mathcal{E}$ in both inequalities leads to

$$V^{\mathcal{E}}(\pi^*) - 2c(n) \leq V^{\mathcal{E}}(\widehat{\pi}_n) \leq V^{\mathcal{E}}(\pi^*) + 2c(n).$$

, where π^* is the policy defined in (4). Finally, we use (D.5) and Theorem 1 to get that

$$\lim_{n\to\infty} \mathbb{E}_D\left[|V^{\mathcal{E}}(\widehat{\pi}_n) - \max_{\pi\in\Pi} V^{\mathcal{E}}(\pi)|\right] \leq \lim_{n\to\infty} \mathbb{E}_D[4c(n)] = 0.$$

This completes the proof of Proposition D.1. □

D.3.3. Proof of Lemma 1

The key argument in the proof of Lemma 1 is a Markov property that we formulate as a lemma below.

Lemma D.1 (Extended Markov Property)**.** *Assume Setting 1. For all subsets* $S \subseteq$

$\{1, \dots, d\}$, *it holds for all* $Z \in \{U^1, \dots, U^p, R\}$ *that*

$$Z \perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S$$
$$\Longrightarrow$$
$$\forall \pi \in \Pi^S : \mathbb{P}^{\pi,e}_{R|X^S} \text{ is the same for all } e \in \mathcal{E},$$

where the symbol $\perp\!\!\!\perp_{\mathcal{G}}$ *denotes d-separation in the graph* $\mathcal{G}$.

Using Lemma D.1, the proof of Lemma 1 goes as follows.

Proof. Let S^{inv} be a d-invariant set and $\pi^{\text{inv}} \in \Pi^{S^{\text{inv}}}$ be a d-invariant policy with respect to S^{inv}. By Definition 3, we have $R \perp\!\!\!\perp_{\mathcal{G}^{S^{\text{inv}}}} e \mid X^{S^{\text{inv}}}$. It then holds by Lemma D.1 for all $x \in \mathcal{X}^{S^{\text{inv}}}$ and all $e, f \in \mathcal{E}$ that

$$\mathbb{E}^{\pi^{\text{inv}},e}\left[R \mid X^{S^{\text{inv}}} = x\right] = \mathbb{E}^{\pi^{\text{inv}},f}\left[R \mid X^{S^{\text{inv}}} = x\right].$$

□

D.3.3.1. Proof of Lemma D.1

Proof. Lemma D.1 corresponds to a global Markov property in the augmented graph (including the non-random environment index). Such results are well-established and used in settings in which $\mathcal{E}$ is finite, for example in influence diagrams Dawid [2002]. The result, however, also holds for more general, even uncountable $\mathcal{E}$.

To prove this, we first fix $S \subseteq \{1, \dots, d\}$, $\pi \in \Pi^S$ and $Z \in \{U, R\}$. Furthermore, let $e \in \mathcal{E}$, let Σ be the discrete σ-algebra on $\mathcal{E}$ and let $\nu_e : \Sigma \to [0, 1]$ be a probability measure that puts non-zero mass on $\{e\}$. We can then replace the environment indicator in the SCM $\mathcal{S}(\pi, e)$ with a random variable E with distribution ν_e. This induces a joint distribution over (E, X, U, A, R) that is globally Markov with respect to the graph $\mathcal{G}^S$, where e is now replaced by E (see Pearl [2009] Thm 1.4.1 or Lauritzen et al. [1990]). Additionally, it satisfies that $(X, U, A, R) \mid E = e$ has the same distribution as the distribution induced by $\mathcal{S}(\pi, e)$. Therefore the d-separation $Z \perp\!\!\!\perp_{\mathcal{G}^S} E \mid X^S$ (which is implied by $Z \perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S$) implies that the joint distribution (E, X, U, A, R) satisfies the following conditional independence

$$Z \perp\!\!\!\perp E \mid X^S. \tag{D.9}$$

Next, denote by p^π the density of (E, X, U, A, R) with respect to a product measure with the discrete measure as the E-component and for all $e \in \mathcal{E}$ denote by $p^{\pi,e}$ the induced density of $\mathcal{S}(\pi, e)$. Then, by construction of the densities and using the conditional independence in (D.9) it holds that for all $x \in \boldsymbol{\mathcal{X}}^S$, all $z \in \text{supp}(Z)$ and all $f \in \mathcal{E}$ with

$\nu_e(f) > 0$ that

$$\begin{aligned} p^{\pi,f}(z \mid X^S = x) &= p^{\pi}(z \mid X^S = x, E = f) \\ &= p^{\pi}(z \mid X^S = x) \\ &=: w_z(x), \end{aligned}$$

The function w_z therefore no longer depends on the environment f nor on ν_e. Since $\nu_e(e) > 0$, this in particular implies that for all $x \in \boldsymbol{\mathcal{X}}^S$ and all $z \in \text{supp}(Z)$ it holds that

$$p^{\pi,e}(z \mid X^S = x) = w_z(x).$$

As this construction works for all $e \in \mathcal{E}$, this completes the proof of Lemma D.1. □

D.3.4. Stable Blanket and Invariance

In this section, if not explicitly stated otherwise, all causal relations such as parents, descendants, ancestors etc. refer to the graph $\mathcal{G}$. Moreover, we use the convention that $k \in \text{DE}(X^k)$, where $\text{DE}(X^k) \subseteq \{1, \dots, d\}$ denotes only the X-variable descendants of X^k. We first define the strongly non-d-invariant set:

$$S_{\text{SNI}} \coloneqq \{j \in \{1, \dots, d\} \mid \exists k \in \text{CI} : j \in \text{DE}(X^k)\},$$

where CI are confounded and directly intervened on nodes (i.e., for $k \in \text{CI}$ there exists $\ell \in \{1, \dots, p\}$ such that $e \to X^k \leftarrow U^\ell \to R$ in $\mathcal{G}$) and define $S_{\text{I}} \coloneqq \{1, \dots, d\} \setminus S_{\text{SNI}}$. Furthermore, we define $S_R \subseteq \{1, \dots, d\}$ to be the set of X-variables such that $j \in S_R$ if and only if $X^j \to R$ in $\mathcal{G}$ or that there exists $\ell \in \{1, \dots, p\}$ such that $X^j \to U^\ell \to R$ in $\mathcal{G}$. The following Lemma will serve as a basis for our proofs of Proposition 1 and Theorem 2.

Lemma D.2 (properties of S_{I})**.** *Assume Setting 1 and Assumption 1. Then, for all $S \in \mathbf{S}_{\text{inv}}$, it holds that $S \subseteq S_{\text{I}}$ and if a d-invariant set exists, it holds that $S_R \subseteq S_{\text{I}}$, S_{I} is d-invariant and*

$$j \in S_{\text{SNI}} \quad \iff \quad X^j \textit{ is strongly non-}d\textit{-invariant.}$$

Proof. The proof is divided into four parts (S.1, S.2, S.3 and S.4):

S.1 We prove that if $S \in \mathbf{S}_{\text{inv}}$ then $S \subseteq S_{\text{I}}$ by contraposition. Let $S \subseteq \{1, \dots, d\}$ be a subset such that there exists $j \in S$ but $j \in S_{\text{SNI}}$. This implies that there exist $k \in \{1, \dots, d\}$ and $\ell \in \{1, \dots, p\}$ such that $e \to X^k \leftarrow U^\ell \to R$ in $\mathcal{G}$ and $j \in \text{DE}(X^k)$. Since $j \in \text{DE}(X^k)$, the path $e \to X^k \leftarrow U^\ell \to R$ is open given X^S, and therefore $R \not\perp\!\!\!\perp_{\mathcal{G}} e \mid X^S$. By Definition 3, this implies that S is not d-invariant, leading to a contradiction.

S.2 In this step, we prove that if a d-invariant set exists, it holds that $S_R \subseteq S_{\text{I}}$. We prove this by contraposition. Assume that there exists $j \in S_R$ such that

$j \in S_{\text{SNI}}$. This implies that there exist $k \in \{1, \dots, d\}$ and $\ell \in \{1, \dots, p\}$ such that $e \to X^k \leftarrow U^\ell \to R$ in $\mathcal{G}$ and $j \in \text{DE}(X^k)$. Now, we construct a contradiction by showing that this would imply that no d-invariant set exists. Let $S \subseteq \{1, \dots, d\}$ be an arbitrary set. There are two possibilities,

(a) $j \in S$: Using the same argument as in Item **S.1**, we have that S is not a d-invariant set.

(b) $j \notin S$: Since $j \in S_R$ but $j \in S_{\text{SNI}}$ there exists a directed path (using that $j \in \text{DE}(X^k)$)

$$e \to X^k \to \underbrace{\cdots}_{\text{part 1}} \to X^j \to \underbrace{\cdots}_{\text{part 2}} \to R,$$

where part 2 either has length zero or consists only of U-variables (by definition of S_R). The only way this path can be blocked by X^S is if either k, j or one of the variables in part 1 are contained in S. However, if this is the case the path $e \to X^k \leftarrow U^\ell \to R$ is open given X^S. Since the edges from X to A are not relevant in this case, this in particular means that $R \not\perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S$, which by Definition 3 implies that S is not d-invariant.

As these are the only two possibilities, we have shown that no d-invariant set exists, which is a contradiction. Therefore $S_R \subseteq S_{\text{I}}$.

S.3 Now we prove that if a d-invariant set exists, then S_{I} is d-invariant. In this step, all the graphical statements are understood to be taken in $\mathcal{G}^{S_{\text{I}}}$. By Lemma D.1, it suffices to show that any path ρ in $\mathcal{G}^{S_{\text{I}}}$ from e to R is blocked by $X^{S_{\text{I}}}$. Let ρ be an arbitrary path from e to R in $\mathcal{G}^{S_{\text{I}}}$. First, we consider the case that ρ enters R through A, i.e., that it has the form

$$e \to \cdots X^j \to A \to R.$$

By construction of $\mathcal{G}^{S_{\text{I}}}$ this path can only be in $\mathcal{G}^{S_{\text{I}}}$ if $j \in S_{\text{I}}$ which implies that it is blocked by $X^{S_{\text{I}}}$. Next, assume that ρ enters R either through a U- or X-variable. Let U^ℓ be the U-variable on ρ that is closest to e and X^j be the X-variable on ρ that is closest to U^ℓ. We consider the two following cases:

(1) U^ℓ does not exist: This implies that ρ does not contain any unobserved variables U and hence ρ can enter R only through an X-variable. By Item **S.2**, we have $S_R \subseteq S_{\text{I}}$ and hence it holds that ρ is blocked by $X^{S_{\text{I}}}$.

(2) U^ℓ exists: ρ has the form

$$\rho: \quad e \to X^r \underbrace{\cdots}_{\text{part 1}} U^\ell \underbrace{\cdots}_{\text{part 2}} \to R,$$

where part 1 could be of length zero or it could consist of further X-variables and part 2 could be of length zero or it could consist of further X-or U-variables. By Assumption 1, we have that there must be an edge from U^ℓ to

R and hence there exists another path

$$\tilde{\rho}: \quad e \to X^r \underbrace{\cdots}_{\text{part 1}} U^\ell \to R,$$

where part 1 corresponds to the part 1 from path ρ. It suffices to show that $\tilde{\rho}$ is blocked by X^{S_I}: whenever $\tilde{\rho}$ is blocked by X^{S_I}, ρ is blocked by X^{S_I} too (as $U^\ell \notin X^{S_\mathrm{I}}$. We now consider the following three cases for $\tilde{\rho}$:

(i) $\tilde{\rho}: \quad e \to \cdots X^j \to U^\ell \to R,$

(ii) $\tilde{\rho}: \quad e \to \cdots \to X^j \leftarrow U^\ell \to R,$

(iii) $\tilde{\rho}: \quad e \to \cdots X^k \leftarrow X^j \leftarrow U^\ell \to R,$

in each case the $\cdots$ can also be of length zero.

Case (i): We show by contradiction that $\tilde{\rho}$ is blocked by X^{S_I}. Assume $\tilde{\rho}$ is open given X^{S_I}. We then have that $j \in S_\mathrm{SNI}$ adjust notation everywhere so that variables are not included but the index, i.e., $j \in S_\mathrm{SNI}$. Let $S \subseteq \{1, \ldots, p\}$ be an arbitrary subset. If $j \in S$, then by the definition of S_SNI there exists $k \in \{1, \ldots, d\}$ and $c \in \{1, \ldots, p\}$ such that $e \to X^k \leftarrow U^c \to R$ in $\mathcal{G}$ and $j \in \mathrm{DE}(X^k)$ and hence $R \not\perp\!\!\!\perp_{\mathcal{G}^{S_\mathrm{I}}} e \mid X^S$. If $j \notin S$, then the path $\tilde{\rho}$ is open given X^S and hence $R \not\perp\!\!\!\perp_{\mathcal{G}^{S_\mathrm{I}}} e \mid X^S$. Therefore, there is no d-invariant set which contradicts to the fact that a d-invariant set exists.

Case (ii): In this case, X^j is a collider on $\tilde{\rho}$. Assume $\tilde{\rho}$ has the form $e \to X^j \leftarrow U^\ell \to R$. This implies that $\mathrm{DE}(X^j) \subseteq S_\mathrm{SNI}$ and hence $\tilde{\rho}$ is blocked by X^{S_I}. Thus, in order for $\tilde{\rho}$ to be open given X^{S_I} it must have the form $e \to \cdots X^k \to X^j \leftarrow U^\ell \to R$. Now, we consider the following two cases separately:

(a) $k \in S_\mathrm{I}$: This directly implies that $\tilde{\rho}$ is blocked by X^{S_I}.

(b) $k \notin S_\mathrm{I}$: By definition of S_I it holds that $\mathrm{DE}(X^k) \cap S_\mathrm{I} = \varnothing$. Hence, also $\mathrm{DE}(X^j) \cap S_\mathrm{I} = \emptyset$ which since X^j is a collider implies that $\tilde{\rho}$ is blocked by X^{S_I}.

We have therefore shown that in Case (ii) the path $\tilde{\rho}$ is blocked by X^{S_I}.

Case (iii): In this case, let X^c be the collider closest to X^j on $\tilde{\rho}$. Again we consider two cases:

(a) $j \in S_\mathrm{I}$: This directly implies that $\tilde{\rho}$ is blocked by X^{S_I}.

(b) $j \notin S_\mathrm{I}$: Since $\mathrm{DE}(X^c) \subseteq \mathrm{DE}(X^j)$, this implies that $\mathrm{DE}(X^c) \cap S_\mathrm{I} = \varnothing$. Hence, the path $\tilde{\rho}$ is blocked by X^{S_I}.

We have therefore shown that in Case (iii) the path $\tilde{\rho}$ is blocked by X^{S_I}. Combining all cases, we have shown that any path $\tilde{\rho}$ from e to R is blocked by X^{S_I} in $\mathcal{G}^{S_\mathrm{I}}$.

S.4 It remains to show that

$$j \in S_\mathrm{SNI} \quad \Leftrightarrow \quad X^j \text{ is strongly non-}d\text{-invariant.}$$

We show each direction separately. First, let $j \in S_{\text{SNI}}$. By the definition of S_{I} it holds that there exists $k \in \{1, \dots, d\}$ and $\ell \in \{1, \dots, p\}$ such that $e \to X^k \leftarrow \cdots \leftarrow U^\ell \to \cdots \to R$ in $\mathcal{G}$ and $j \in \text{DE}(X^k)$. As this path does not involve A it is contained in $\mathcal{G}^S$ for all subsets $S \subseteq \{1, \dots, d\}$. Moreover, since X^k is a collider and the only X-variable on this path, it holds that this path will be open given $X^{S \cup \{j\}}$ for all subsets $S \subseteq \{1, \dots, d\}$. Therefore, X^j is strongly non-d-invariant. Next, to show the reverse direction let $j \in S_{\text{I}}$. Then, by Item **S.3** it holds that S_{I} is d-invariant. So in particular $R \perp\!\!\!\perp_{\mathcal{G}^{S_{\text{I}}}} e \mid X^{S_{\text{I}}}$, which since $j \in S_{\text{I}}$ implies that j is not strongly non-d-invariant.

This completes the proof of Lemma D.2. □

As shown in Lemma D.2 the set S_{I} is d-invariant and contains all d-invariant sets if a d-invariant set exists. It will be used in the proofs of Proposition 1 and Theorem 2 to find the optimal d-invariant policy, as it encodes all invariant available information about the reward. The set S_{I} is strongly related to stable blankets as defined in Pfister et al. [2021].

D.3.5. Proof of Proposition 1

Proof. As before, we define $\pi_a(a' \mid x) := \mathbb{1}\big[a' = a\big]$ as the policy that always selects the action a.

The proof is divided into three steps (S.1, S.2 and S.3)

S.1 In the first step, we use the set S_{I} defined in Appendix D.3.4 to derive an upper bound on the expected reward of arbitrary d-invariant policies. First by the same arguments as in (D.2) we get that, for all $S \in \mathbf{S}_{\text{inv}}$, all $\pi^S \in \Pi^S$, all $x \in \boldsymbol{\mathcal{X}}^S$ and all $e \in \mathcal{E}$, it holds that

$$\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S = x] \geq \mathbb{E}^{\pi^S, e}[R \mid X^S = x].$$

Taking the expectation over X^S on both sides yields

$$\begin{aligned} \mathbb{E}^e \left[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S] \right] &\geq \mathbb{E}^e \left[\mathbb{E}^{\pi^S, e}[R \mid X^S] \right] \\ &= \mathbb{E}^{\pi^S, e}[R]. \end{aligned} \tag{D.10}$$

Next, we make use of the set S_{I} defined in Appendix D.3.4. By Lemma D.2, it holds that $S \subseteq S_{\text{I}}$ for all $S \in \mathbf{S}_{\text{inv}}$. We then have, for all $S \in \mathbf{S}_{\text{inv}}$, $a \in \mathcal{A}$, and $e \in \mathcal{E}$, that

$$\mathbb{E}^{\pi_a, e}[R \mid X^S, X^{S_{\text{I}} \setminus S}] = \mathbb{E}^{\pi_a, e}[R \mid X^{S_{\text{I}}}], \tag{D.11}$$

This is closely related to the predictiveness property of stable blankets (see Pfister et al. [2021]).

Now, we expand the conditional expectation and get for all $e \in \mathcal{E}$ that

$$\begin{aligned}&\mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X^S]\right]\\&= \mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^e_{X^{S_\mathrm{I}\setminus S}}\left[\mathbb{E}^{\pi_a,e}[R \mid X^S, X^{S_\mathrm{I}\setminus S}]\right]\right],\end{aligned}$$

and by Jensen's inequality,

$$\begin{aligned}&\leq \mathbb{E}^e\left[\mathbb{E}^e_{X^{S_\mathrm{I}\setminus S}}\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X^S, X^{S_\mathrm{I}\setminus S}]\right]\right]\\&= \mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X^S, X^{S_\mathrm{I}\setminus S}]\right],\end{aligned}$$

and by (D.11),

$$= \mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X^{S_\mathrm{I}}]\right].$$

Combining this with (D.10), we have

$$\mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X^{S_\mathrm{I}}]\right] \geq \mathbb{E}^{\pi^S,e}[R]. \tag{D.12}$$

S.2 In the second step, we use the upper bound (D.12) to show that, for different environments, the same set of policies is optimal. For all $\pi \in \Pi_{\text{inv}}$ it holds by Lemma D.2 that $\pi \in \Pi^{S_\mathrm{I}}$ and by Lemma 1 that $\mathbb{E}^\pi[R \mid X^{S_\mathrm{I}}]$ does not depend on e (since S_I is d-invariant). Let $\bar{\pi} \in \Pi_{\text{inv}}$ be a policy that satisfies for all $a \in \mathcal{A}$ and for μ-a.e. $x \in \boldsymbol{\mathcal{X}}^{S_\mathrm{I}}$

$$\bar{\pi}(a|x) > 0 \implies a \in \arg\max_{a'\in\mathcal{A}} \mathbb{E}^{\pi_{a'}}\left[R \mid X^{S_\mathrm{I}} = x\right]. \tag{D.13}$$

Then, it holds for all $e \in \mathcal{E}$ that

$$\begin{aligned}\mathbb{E}^{\bar{\pi},e}[R] &= \mathbb{E}^e\left[\mathbb{E}^{\bar{\pi}}[R \mid X^{S_\mathrm{I}}]\right]\\&= \mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}[R \mid X^{S_\mathrm{I}}]\right].\end{aligned}$$

Since (D.12) holds for all $S \in \mathbf{S}_{\text{inv}}$, this directly implies that $\bar{\pi} \in \arg\max_{\pi\in\Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R]$. We now show the reverse direction, i.e., if $\pi^* \in \arg\max_{\pi\in\Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R]$, then π^* satisfies (D.13). Let $\pi^* \in \arg\max_{\pi\in\Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R]$. By (D.12), we have, for all $e \in \mathcal{E}$,

$$\mathbb{E}^{\pi^*,e}[R] = \mathbb{E}^e\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}[R \mid X^{S_\mathrm{I}}]\right]. \tag{D.14}$$

Since, for all $e \in \mathcal{E}$, the distribution of X has full support (by the assumption in Setting 1), π^* satisfies for all $a \in \mathcal{A}$ and for μ-a.e. $x \in \boldsymbol{\mathcal{X}}^{S_\mathrm{I}}$

$$\pi^*(a|x) > 0 \implies a \in \arg\max_{a'\in\mathcal{A}} \mathbb{E}^{\pi_{a'}}\left[R \mid X^{S_\mathrm{I}} = x\right]. \tag{D.15}$$

Thus, π^* satisfies (D.15) if and only if $\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R]$. Furthermore, since (D.15) does not depend on e, it then holds for all $e, f \in \mathcal{E}$ that

$$\arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R] = \arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,f}[R]. \tag{D.16}$$

S.3 In the third step, we are now ready to prove the main result of the proposition. To this end, let

$$\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \sum_{e \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi,e}[R].$$

Then, from (D.16) we have for all $e, f \in \mathcal{E}$ that

$$\arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R] = \arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,f}[R].$$

So in particular, for all $e \in \mathcal{E}$, it holds that

$$\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \mathbb{E}^{\pi,e}[R].$$

Thus it holds for all $e \in \mathcal{E}$, all $S \in \mathbf{S}_{\text{inv}}$ and all $\pi^S \in \Pi^S$ that

$$\mathbb{E}^{\pi^*,e}[R] \geq \mathbb{E}^{\pi^S,e}[R].$$

Taking the infimum over $e \in \mathcal{E}$ on both sides yields

$$V^{\mathcal{E}}(\pi^*) = \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi^*,e}[R] \geq \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi^S,e}[R] = V^{\mathcal{E}}(\pi^S).$$

Because this inequality holds for all $S \in \mathbf{S}_{\text{inv}}$ and all $\pi^S \in \Pi^S$, this implies

$$\forall \pi \in \Pi_{\text{inv}} : \quad V^{\mathcal{E}}(\pi^*) \geq V^{\mathcal{E}}(\pi). \tag{D.17}$$

This completes the proof of Proposition 1.

□

D.3.6. Proof of Theorem 2

Proof. We first prove the first statement of Theorem 2. Fix a policy

$$\pi^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} \sum_{e \in \mathcal{E}^{\text{obs}}} \mathbb{E}^{\pi,e}[R].$$

Using the same argument as we made in Appendix D.3.5(**S.3**), we get that for all $e \in \mathcal{E}$ it holds that

$$\mathbb{E}^{\pi^*,e}[R] = \mathbb{E}^e \left[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a}[R \mid X^{S_{\text{I}}}] \right]. \tag{D.18}$$

Hence by Jensen's inequality it holds that

$$\begin{aligned}\mathbb{E}^{\pi^*,e}[R] &\geq \max_{a\in\mathcal{A}} \mathbb{E}^{e}\left[\mathbb{E}^{\pi_a}[R \mid X^{S_{\mathrm{I}}}]\right] \\ &= \max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R].\end{aligned}$$

This completes the proof of the first statement.

Next, we prove the second statement of Theorem 2. To do so, we use the following lemma, which is proved in Appendix D.3.6.1 below.

Lemma D.3 (Upper bound)**.** *Assume Setting 1, Assumptions 1, 3 and 2, and that* $\mathbf{S}_{\mathrm{inv}} \neq \varnothing$*. Let* $\pi \in \Pi \setminus \Pi_{\mathrm{inv}}$ *be an arbitrary non-d-invariant policy. Then it holds that*

$$V^{\mathcal{E}}(\pi) \leq \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_{\mathrm{I}}}}\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}\left[R \mid X^{S_{\mathrm{I}}}\right]\right].$$

To finish the proof of Theorem 2, fix again a policy

$$\pi^* \in \arg\max_{\pi\in\Pi_{\mathrm{inv}}} \sum_{e\in\mathcal{E}^{\mathrm{obs}}} \mathbb{E}^{\pi,e}[R].$$

Then, by Proposition 1, it holds that

$$\forall \pi \in \Pi_{\mathrm{inv}} : \quad V^{\mathcal{E}}(\pi) \leq V^{\mathcal{E}}(\pi^*). \tag{D.19}$$

Furthermore, Lemma D.3 together with (D.18) implies that

$$\forall \pi \in \Pi \setminus \Pi_{\mathrm{inv}} : \quad V^{\mathcal{E}}(\pi) \leq V^{\mathcal{E}}(\pi^*). \tag{D.20}$$

Combining (D.19) and (D.20) concludes the proof of Theorem 2. □

D.3.6.1. Proof of Lemma D.3

Proof. Recall the terminology and notation from Appendix D.3.4. The proof can be split into two parts:

1. We first prove that if $e \in \mathcal{E}$ is a confounding removing environment it holds for all $\pi \in \Pi$ that
$$\forall j \in S_{\mathrm{SNI}} : R \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} X^j \mid X^{S_{\mathrm{I}}}, A. \tag{D.21}$$
2. We then prove the upper bound using step 1) as the main argument.

Step 1) Let $e \in \mathcal{E}$ be a confounding removing environment and fix $j \in S_{\mathrm{SNI}}$ and $\pi \in \Pi$. By Lemma D.2 it holds that X^j is strongly non-d-invariant. Therefore, since e is a confounding removing environment, we get that

$$X^j \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} U. \tag{D.22}$$

Now, let ρ be an arbitrary path from X^j to R in $\mathcal{G}^{\pi,e}$. We consider the following (separate) cases that can occur:

(a) ρ enters R through A: Then the path ρ is blocked by X^{S_I} and A because A is not a collider and hence blocks ρ.

(b) ρ only contains A and X-variables and enter R through X-variables: Then there exists $k \in \{1, \dots, d\}$ such that ρ ends with $X^k \to R$. This implies that $k \in S_R$ since $\mathcal{G}^{\pi,e}$ is a sub-graph of $\mathcal{G}$. Furthermore, since by Lemma D.2 (recall that $\mathbf{S}_\text{inv} \neq \varnothing$) $S_R \subseteq S_\text{I}$, this implies that $k \in S_\text{I}$. Hence, ρ is blocked by X^{S_I} and A because X^k is not a collider.

(c) ρ contains at least one U-variable: Let $\ell \in \{1, \dots, p\}$ such that U^ℓ is the U-variable closest to X^j on ρ, i.e., ρ has the form

$$\underbrace{X^j \cdots U^\ell}_{\gamma} \cdots \to R.$$

Now, by (D.22) it holds that γ is blocked (given the empty set) in $\mathcal{G}^{\pi,e}$ and by construction it only consists of X-variables (except U^ℓ). Therefore, there must be at least one collider on γ. Let X^k be the collider closest to U^ℓ and let X^m (this could be X^j) the variable that comes right before X^k on γ, i.e.,

$$X^j \cdots X^m \to X^k \leftarrow \cdots U^\ell.$$

We consider two cases:

(i) First, assume that $\text{DE}(X^k) \cap S_\text{I} \neq \varnothing$ (in $\mathcal{G}^{\pi,e}$), then it holds, by the definition of S_I and since $\mathcal{G}^{\pi,e}$ is a subgraph of $\mathcal{G}$, that $m \in S_\text{I}$ as well (otherwise none of the descendants of X^k could be in S_I as $\text{DE}(X^k) \subset \text{DE}(X^m)$). However, X^m is not a collider and therefore ρ is blocked given X^{S_I} and A.

(ii) Second, assume $\text{DE}(X^k) \cap S_\text{I} = \varnothing$, then it in particular holds that $k \in S_\text{SNI}$ which by Lemma D.2 implies that X^k is strongly non-d-invariant. Hence, because e is a confounding removing environment, it holds that $X^k \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} U$. However, X^k was selected to be the collider closest to U^ℓ which means that the part of γ from X^k to U^ℓ is open in $\mathcal{G}^{\pi,e}$ leading to a contradiction.

We have therefore shown that the path ρ is always blocked given X^{S_I} and A. Since ρ was arbitrary this implies that $R \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} X^j \mid X^{S_\text{I}}, A$.

Step 2)

Now, we are ready to prove the main result. Let $\pi \in \Pi \setminus \Pi_\text{inv}$ be an arbitrary non-d-invariant policy, and let $S \subseteq \{1, \dots, d\}$ such that $\pi \in \Pi^S$. We have

$$\begin{aligned} &V^{\mathcal{E}}(\pi) \\ &= \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi,e}\left[R\right], \end{aligned}$$

by the tower property of conditional expectation,

$$= \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\mathbb{E}^{\pi,e}\left[R \mid X^{S_\mathrm{I}}, X^{S\setminus S_\mathrm{I}}\right]\right]$$
$$= \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\int \mathbb{E}^{\pi_a,e}\left[R \mid X^{S_\mathrm{I}}, X^{S\setminus S_\mathrm{I}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\right].$$

Now, we use Assumption 2. For each $e \in \mathcal{E}$ we choose a confounding removing environment $f(e)$ such that $\mathbb{P}^{\pi,f(e)}_X = \mathbb{P}^{\pi,e}_X$. Because the confounding removing environments are a subset of $\mathcal{E}$, we have

$$V^{\mathcal{E}}(\pi)$$
$$= \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\int \mathbb{E}^{\pi_a,e}\left[R \mid X^{S_\mathrm{I}}, X^{S\setminus S_\mathrm{I}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\right]$$
$$\leq \inf_{e\in\mathcal{E}} \mathbb{E}^{f(e)}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\int \mathbb{E}^{\pi_a,f(e)}\left[R \mid X^{S_\mathrm{I}}, X^{S\setminus S_\mathrm{I}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\right].$$

Using that $\mathbb{P}^{\pi,f(e)}_X = \mathbb{P}^{\pi,e}_X$, we then have

$$V^{\mathcal{E}}(\pi)$$
$$\leq \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\int \mathbb{E}^{\pi_a,f(e)}\left[R \mid X^{S_\mathrm{I}}, X^{S\setminus S_\mathrm{I}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\right].$$

Next, we use (D.21) which states that for all $j \in \{1,\dots,d\}$ it holds that $R \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} X^j \mid X^{S_\mathrm{I}}, A$. Then, by the Markov property, we get

$$V^{\mathcal{E}}(\pi)$$
$$\leq \inf_{e\in\mathcal{E}} \mathbb{E}^{e}_{X^{S_\mathrm{I}},X^{S\setminus S_\mathrm{I}}} \left[\int \mathbb{E}^{\pi_a,f(e)}\left[R \mid X^{S_\mathrm{I}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\right],$$

we can then omit $f(e)$ since S_{I} is a d-invariant set (by Lemma D.2 since $\mathbf{S}_{\mathrm{inv}} \neq \varnothing$),

$$
\begin{aligned}
&= \inf_{e \in \mathcal{E}} \mathbb{E}^e_{X^{S_{\mathrm{I}}}, X^{S \setminus S_{\mathrm{I}}}} \Bigg[\int \mathbb{E}^{\pi_a} \left[R \mid X^{S_{\mathrm{I}}} \right] \\
&\qquad\qquad \pi(a|X^S)\, \mu(\mathrm{d}a) \Bigg] \\
&= \inf_{e \in \mathcal{E}} \mathbb{E}^e_{X^{S_{\mathrm{I}}}} \Bigg[\int \mathbb{E}^{\pi_a} \left[R \mid X^{S_{\mathrm{I}}} \right] \\
&\qquad \mathbb{E}^e_{X^{S \setminus S_{\mathrm{I}}}} \left[\pi(a|X^S) \right] \mu(\mathrm{d}a) \Bigg],
\end{aligned}
$$

letting $\tilde{\pi}(a|X^{S_{\mathrm{I}}}) \coloneqq \mathbb{E}^e_{X^{S \setminus S_{\mathrm{I}}}}[\pi(a|X^S)]$,

$$
\begin{aligned}
&= \inf_{e \in \mathcal{E}} \mathbb{E}^e_{X^{S_{\mathrm{I}}}} \left[\int \mathbb{E}^{\pi_a} \left[R \mid X^{S_{\mathrm{I}}} \right] \tilde{\pi}(a \mid X^{S_{\mathrm{I}}})\, \mu(\mathrm{d}a) \right] \\
&\leq \inf_{e \in \mathcal{E}} \mathbb{E}^e_{X^{S_{\mathrm{I}}}} \left[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a} \left[R \mid X^{S_{\mathrm{I}}} \right] \right].
\end{aligned}
$$

□

D.3.7. Proof of Proposition 2

Proof. Fix a set $S \subseteq \{1, \ldots, p\}$, and let $\pi, \tilde{\pi} \in \Pi^S$. Assume $H_0(S, \pi, \mathcal{E})$ is true. By Item 3(i), we have that $R \perp\!\!\!\perp_{\mathcal{G}^S} e \mid X^S$. Furthermore, since $\tilde{\pi} \in \Pi^S$ this implies by Lemma D.1 that $\mathbb{P}^{\tilde{\pi}, e}_{R|X^{S^{\mathrm{inv}}}}$ is the same for all $e \in \mathcal{E}$ which implies that $H_0(S, \tilde{\pi}, \mathcal{E})$ is true. This concludes the proof of Proposition 2. □

D.3.8. Proof of Proposition 3

Proof. Let $S^* \coloneqq \mathrm{AN}(R)$ be the set of observed ancestors of R. In this proof, all the graphical statements are understood to be taken in $\mathcal{G}^{S^*}$. Assume that a d-invariant set S exists. Then, by Theorem 2 of Tian et al. [1998], $S \cap S^*$ is d-invariant, too (indeed, S intersected with all ancestors of R is d-invariant but as S does not contain any hidden variables, this set equals $S \cap S^*$).

We are now ready to prove the statement of the proposition. From (14) we have,

$$
\begin{aligned}
&\liminf_{n \to \infty} \mathbb{P}(\hat{S}^n_{\mathrm{AN}} \subseteq S^*) \\
=&\liminf_{n \to \infty} \mathbb{P}\Big(\bigcap_{S : \psi^S(D^{e_1, \pi^S}, \ldots, D^{e_L, \pi^S}) = 1} S \subseteq S^* \Big) \\
\geq&\liminf_{n \to \infty} \mathbb{P}(\psi^{S \cap S^*}(D^{e_1, \pi^{S \cap S^*}}, \ldots, D^{e_L, \pi^{S \cap S^*}}) = 1) \\
\geq& 1 - \alpha,
\end{aligned}
$$

where the last inequality follows by Proposition D.2. This completes the proof of Proposition 3. □

D.4. Connection to Random Environments

It is possible to define multi-environment contextual bandits using random environments.

Setting D.1 (Random Environment Contextual Bandits)**.** *Let* $X = (X^1, \dots, X^d) \in \boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \dots \times \mathcal{X}^d$, $U = (U^1, \dots, U^p) \in \boldsymbol{\mathcal{U}} = \mathcal{U}^1 \times \dots \times \mathcal{U}^p$, $A \in \mathcal{A} = \{a^1, \dots, a^k\}$, $R \in \mathbb{R}$, $E \in \mathcal{E}$. *For any* $\pi \in \{\pi : \boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})\}$, *let* g_π *denote the function that ensures, for all* $x \in \boldsymbol{\mathcal{X}}$, $g_\pi(x, \varepsilon_A)$ *equals* $\pi(x)$ *in distribution for a uniformly distributed* ε_A. *Now, consider functions* s, h, *and* f, *a factorizing distribution* $\mathbb{P}_\varepsilon = \mathbb{P}_{\varepsilon_E} \times \mathbb{P}_{\varepsilon_U} \times \mathbb{P}_{\varepsilon_X} \times \mathbb{P}_{\varepsilon_A} \times \mathbb{P}_{\varepsilon_R}$ *whose* ε_A *component is uniform, and a structural causal model* $\mathcal{S}(\pi)$ *given by*

$$\mathcal{S}(\pi): \quad \begin{cases} E \coloneqq \varepsilon_E \\ U \coloneqq s(X, \varepsilon_U) \\ X \coloneqq h(X, U, E, \varepsilon_X) \\ A \coloneqq g_\pi(X, \varepsilon_A) \\ R \coloneqq f(X, U, A, \varepsilon_R). \end{cases}$$

Assume further that for all π, *the SCM induces a unique distribution over* (E, X, U, A, R), *which we denote by* $\mathbb{P}^\pi$. *The structure of the SCM* $\mathcal{S}(\pi, e)$ *can be also visualized by a graph* $\mathcal{G}$ *which is constructed in a similar way to the graph in Setting 1, except that the environment becomes one of the variable nodes in this graph.*

Remark D.1. Setting D.1 is a special case of Setting 1 in the following sense: Assume, starting from Setting D.1, for all $i \in \{1, \dots, n\}$ that $(X_i, U_i, A_i, R_i, E_i)$, are independent and distributed according to $\mathbb{P}^{\pi_i}_{X,U,A,R,E}$. Then, defining $h_e(\cdot, \cdot) \coloneqq h(\cdot, e, \cdot)$, we have that, for all $i \in \{1, \dots, n\}$, (X_i, U_i, A_i, R_i), are independent and distributed according to $\mathbb{P}^{\pi_i, E_i}_{X,U,A,R}$, using Setting 1.

D.5. Details for Section 4.2

In Section 4.2, we propose to use the resampling procedure in Thams et al. [2021] to test the hypothesis of invariance under a test policy $\pi^S \in \Pi^S$.

For every $e \in \mathcal{E}^{\text{obs}}$, we have a data set D^e consisting of n_e observations $D^e_i = (X^e_i, A^e_i, R^e_i, \pi^0(A^e_i|X^e_i))$ is available.[1] For all $e \in \mathcal{E}^{\text{obs}}$ and all $i \in \{1, \dots, n_e\}$ define the relative weights as

$$r(D^e_i) \coloneqq \frac{\pi^S(A^e_i|X^e_i)}{\pi^0(A^e_i|X^e_i)}. \tag{D.23}$$

[1] It is possible to allow for a different initial policy π^0_i at each observation i. One then needs to define the relative weights $r(D^e_i) \coloneqq \pi^S(A^e_i|X^e_i)/\pi^0_i(A^e_i|X^e_i)$.

Then, for all $e \in \mathcal{E}^{\text{obs}}$, we draw a weighted resample $D^{e,\pi^S} := (D^e_{i_1}, \dots, D^e_{i_{m_e}})$ of size m_e from D^e with weights

$$w^e_{i_1,\dots,i_{m_e}} := \begin{cases} \frac{\prod_{\ell=1}^{m_e} r(D^e_{i_\ell})}{\sum_{(j_1,\dots,j_{m_e}) \text{ distinct}} \prod_{\ell=1}^{m_e} r(D^e_{j_\ell})} & (i_1, \dots, i_{m_e}) \text{ distinct} \\ 0 & \text{otherwise.} \end{cases} \tag{D.24}$$

We then apply an invariance test to the resampled data $D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S}$. A family of invariance tests $\{\varphi^S\}_{S \subseteq \{1,\dots,d\}}$ is a collection of functions such that for each S, φ^S is a function (into $\{0,1\}$) that takes data from environments $e_1, \dots, e_L$, each of size m_{e_i}, and tests whether S is invariant. Here, $\varphi^S = 1$ indicates that we reject the hypothesis of invariance. We say the test has pointwise asymptotic level if for all invariant sets S and all $\pi \in \Pi^S$ it holds that

$$\limsup_{\min\{m_{e_1},\dots,m_{e_L}\} \to \infty} \mathbb{P}^\pi(\varphi^S(D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S}) = 1) \leq \alpha.$$

We state that the overall procedure (resampling and then testing) has asymptotic level as long as the test φ^S has asymptotic level. For simplicity, we assume that $n_{e_1} = \dots = n_{e_L} =: n$ and $m_{e_1} = \dots = m_{e_L} =: m$. The following result follows directly from [Thams et al., 2021, Theorem 1]

Proposition D.2. *Let $S \subseteq \{1, \dots, d\}$ and suppose that for each environment $e_1, \dots, e_L$, we observe a data set D^e consisting of n observations $D^e_i = (X^e_i, A^e_i, R^e_i, \pi^0(A^e_i \mid X^e_i))$. Consider $\pi^S \in \Pi^S$ and assume that for all $e \in \mathcal{E}$, $\mathbb{E}^{\pi^0}[r(D^e_i)^2] < \infty$, where r is defined in (D.23). Let $m = o(\sqrt{n})$ and for all e, let $D^{e,\pi^S} := (D^e_{i_1}, \dots, D^e_{i_m})$ be a resample of D^e drawn with weights given by (D.24). Let φ^S be a hypothesis test for invariance of the conditional expectation $\mathbb{E}^{\pi^S,e}[R \mid X^S]$ that has pointwise asymptotic level $\alpha \in (0,1)$ when φ^S is applied to data sampled with π^S. Applying φ^S to the resampled data yields pointwise asymptotic level, that is,*

$$\limsup_{n \to \infty} \mathbb{P}^{\pi^0}(\varphi^S(D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S}) = 1) \leq \alpha$$

if S is invariant.

Proof. We only show that this problem with environments can be cast in the setting of Thams et al. [2021], which has no reference to environments. Here, we assume that we have the same number of observations in each environment. The main idea is to create a data set $D^{\mathcal{E}}$, such that each observation in $D^{\mathcal{E}}$ consists of an observation from each of the environments D^e.

First, we randomly permute the observations within each data set D^e to obtain a set $\tilde{D}^e$. Then, we construct an auxiliary data set $D^{\mathcal{E}}$, where the i'th observation $D^{\mathcal{E}}_i$ of $D^{\mathcal{E}}$ is the concatenation of the i'th observation (after permutation) from each of the environments, $D^{\mathcal{E}}_i := (\tilde{D}^{e_1}_i, \dots, \tilde{D}^{e_L}_i)$.

We can now apply the resampling methodology from Thams et al. [2021] to draw a sequence $(D^{\mathcal{E}}_{i_1}, \dots, D^{\mathcal{E}}_{i_m})$ with weights given by

$$w^{\mathcal{E}}_{i_1,\dots,i_m} := \begin{cases} \frac{\prod_{\ell=1}^m r(D^{\mathcal{E}}_{i_\ell})}{\sum_{(j_1,\dots,j_m) \text{ distinct}} \prod_{\ell=1}^m r(D^{\mathcal{E}}_{j_\ell})} & (i_1, \dots, i_m) \text{ distinct} \\ 0 & \text{otherwise.} \end{cases}$$

where

$$r(D^{\mathcal{E}}_i) := \frac{\pi^S(\tilde{A}^{e_1}_i \mid \tilde{X}^{e_1}_i)}{\pi^0(\tilde{A}^{e_1}_i \mid \tilde{X}^{e_1}_i)} \cdots \frac{\pi^S(\tilde{A}^{e_L}_i \mid \tilde{X}^{e_L}_i)}{\pi^0(\tilde{A}^{e_L}_i \mid \tilde{X}^{e_L}_i)},$$

and $\tilde{X}^e_i, \tilde{A}^e_i$ are the i'th observation of $\tilde{D}^e$. Because the observations are independent, both within and between environments, the probability of drawing the resampled data set $(D^{\mathcal{E}}_{i_1}, \dots, D^{\mathcal{E}}_{i_m}) = ((D^{e_1}_{i_1}, \dots, D^{e_L}_{i_1}), \dots, (D^{e_1}_{i_m}, \dots, D^{e_L}_{i_m}))$ is equal to the probability of drawing first m observations from e_1, $(D^{e_1}_{i_1}, \dots, D^{e_1}_{i_m})$, and then m from e_2 etc. The result then follows directly from Thams et al. [2021]. □

In other words, we can test whether S is invariant by resampling the data and applying an invariance test on the resampled data set. Proposition D.2 states that this procedure holds level asymptotically. We assume knowledge of the initial policy π^0 to ease our presentation. We can, in fact, show the pointwise asymptotic validity even if the initial policy π^0 is unknown and has to be estimated from the offline data (see Thams et al. [2021] Theorem 2).

D.6. Algorithm: Off-policy Invariant Causal Prediction

Below, we present an algorithm for finding the causal ancestors $\text{AN}(R)$ of the reward R under a change in policy.

Algorithm D.1 Off-policy Invariant Causal Prediction

Input data $D = (D^{e_1}, \dots, D^{e_L})$, test function `pv`, initial policy π_0, resampling size $\mathbf{m} := (m_1, \dots, m_L) = (\sqrt{|D^{e_1}|}, \dots, \sqrt{|D^{e_L}|})$

initialize the collection of invariant sets $\mathbf{S}_{\text{inv}} \leftarrow \{\}$ ▷ loop over all subsets
for $S \in \mathcal{P}(\{1, \dots, d\})$ **do** ▷ test for invariance
  if $\pi^S \neq$ `null` **then**
    `is_inv` $\leftarrow$ `test_inv`$(D, \pi^S, \text{pv}, S, \mathbf{m})$ ▷ (see Algorithm 2)
  else `is_inv` $\leftarrow$ `test_inv_opt`$_\pi(D, \text{pv}, S, \mathbf{m})$ ▷ (see Algorithm D.2 in Appendix D.8)
  if `is_inv` **then** add S to $\mathbf{S}_{\text{inv}}$ ▷ update the accepted invariant set
$\hat{S}_{\text{AN}} \leftarrow \bigcap_i \mathbf{S}_{\text{inv}}[i]$ ▷ get the estimated causal ancestors
Output: the estimated causal ancestors $\hat{S}_{\text{AN}}$

D.7. Faster power optimization

In Section 4.5.2, we show that we can optimize the power to detect non-invarince by gradient descent. In particular, the gradient is

$$\nabla J(\theta) = \mathbb{E}\left[\nabla \log \mathbb{P}(D^{\pi_\theta^S} \mid D)\mathtt{pv}(D^{\pi_\theta^S})\right],$$

where $D^{\pi_\theta^S}$ is a resample of the data D and $\mathtt{pv}$ is a function returning a p-value of our invariance test. $\mathbb{P}(D^{\pi_\theta^S} \mid D)$ is given by (D.24), but as discussed in Thams et al. [2021], this may be infeasible to compute if n is very large.

As a computationally efficient alternative, Thams et al. [2021] proposes an approximate resampling scheme, where a sequence $(i_1, \dots, i_{m_e})$ (distinct or non-distinct) is sampled with replacement. That is, the weights are given by

$$\begin{aligned} w_{\theta,(i_1,\dots,i_{m_e})} &:= \frac{\prod_{\ell=1}^{m_e} r_\theta(D^e_{i_\ell})}{\sum_{(j_1,\dots,j_m)} \prod_{\ell=1}^{m_e} r_\theta(D^e_{j_\ell})} \\ &= \frac{\prod_{\ell=1}^{m_e} r_\theta(D^e_{i_\ell})}{\left(\sum_{j=1}^{n_e} r_\theta(D^e_j)\right)^{m_e}}. \end{aligned}$$

This expression is much easier to compute than (D.24), because the denominator is a sum over n_e terms (instead of $n_e!/(n_e - m_e)!$). In particular, we get

$$\begin{aligned} \nabla_\theta \log \mathbb{P}(D^{\pi_\theta^S} \mid D) &= \nabla_\theta \log w_{\theta,(i_1,\dots,i_{m_e})} \\ &= \sum_{\ell=1}^{m_e} \nabla_\theta \log r_\theta(D^e_{i_\ell}) - m_e \nabla_\theta \log \sum_{j=1}^{n_e} r_\theta(D^e_j). \end{aligned}$$

Algorithm D.2 splits the data in two halves: we optimize power on the first half of the data and test for invariance on the second half. We only use the above approximation for the power optimization, where we need to explicitly compute the normalization constant of the weights. In the second half of Algorithm D.2, we use (D.24) (i.e., we do not use the approximate weights), because Proposition D.2 requires the weights to be those given in (D.24). If n is so large that we cannot sample by explicitly computing the weights (D.24), there are several options for sampling from the scheme without computing the denominator – see Thams et al. [2021] for a variety of approaches.

D.8. Invariance test with optimized test policy

In this section we provide Algorithm D.2, which tests the invariance of a set by choosing a test policy π^S that optimizes the power of the invariance test, as discussed in Section 4.5.2.

Algorithm D.2 Testing the invariance of a set S with optimization over test policies π^S

`test_inv_opt_`π(data $D = (D^{e_1}, \dots, D^{e_L})$, function `pv` yielding the p-value of an invariance test, target set S, resampling size $(m_1, \dots, m_L) = (\sqrt{|D^{e_1}|/2}, \dots, \sqrt{|D^{e_L}|/2})$, learning rate γ=1e-3, significance level α

▷ sample splitting

for $e = e_1, \dots, e_L$ **do**
  $n_{e,sp} \leftarrow \text{ceil}(|D^e|/2)$
  $D^{e,1} \leftarrow \{(x_i^e, a_i^e, r_i^e, \pi^0(a_i^e|x_i^e))\}_{i=1}^{n_{e,sp}}$
  $D^{e,2} \leftarrow \{(x_i^e, a_i^e, r_i^e, \pi^0(a_i^e|x_i^e))\}_{i=n_{e,sp}+1}^{|D^e|}$

▷ optimizing power

Initialize policy parameters θ
while not converged **do**
  for $e = e_1, \dots, e_L$ **do**
    for $i = 1$ to $n_{e,sp}$ **do**
      compute weights: $r_i^e \leftarrow \dfrac{\pi_\theta^S(a_i^e \mid x_i^{e,S})}{\pi^0(a_i^e \mid x_i^e)}$
    draw $D^{e,\pi_\theta^S} \coloneqq (D_{i_1}^{e,1}, \dots, D_{i_{m_e}}^{e,1})$ with replacement from $D^{e,1}$ with probabilities $\propto r_i^e$
  $D^{1,\pi_\theta^S} \leftarrow (D^{e_1,\pi_\theta^S}, \dots, D^{e_L,\pi_\theta^S})$
  compute p-value: $\texttt{pv}(D^{1,\pi_\theta^S})$
  compute gradient: $\nabla \log \mathbb{P}(D^{1,\pi_\theta^S})$
  update policy parameters: $\theta \leftarrow \theta - \gamma \texttt{pv}(D^{1,\pi_\theta^S}) \nabla \log \mathbb{P}(D^{1,\pi_\theta^S})$

▷ verifying invariance condition

for $e = e_1, \dots, e_L$ **do**
  for $i = n_{e,sp} + 1$ to $|D^e|$ **do**
    compute weights: $r_i^e \leftarrow \dfrac{\pi_\theta^S(a_i^e \mid x_i^{e,S})}{\pi^0(a_i^e \mid x_i^e)}$
  draw $D^{e,\pi_\theta^S} \coloneqq (D_{i_1}^{e,2}, \dots, D_{i_{m_e}}^{e,2})$ with replacement from $D^{e,2}$ with probabilities $\propto r_i^e$
$D^{2,\pi_\theta^S} \leftarrow (D^{e_1,\pi_\theta^S}, \dots, D^{e_L,\pi_\theta^S})$
`is_invariant` $\leftarrow \texttt{pv}(D^{2,\pi_\theta^S}) \geq \alpha$
Return: `is_invariant`

D.9. Simulation Details

D.9.1. Data Generating Process

We generate data from the following SCM $\mathcal{S}(\pi, e)$:

$$U \coloneqq \varepsilon_U, \quad X^1 \coloneqq \gamma_e U + \varepsilon_{X^1}, \quad X^2 \coloneqq \alpha_e + \varepsilon_{X^2},$$
$$A \sim \pi(A \mid X^1, X^2), \quad R \coloneqq \beta_{A,1} X^2 + \beta_{A,2} U + \varepsilon_R,$$

where $\varepsilon_U, \varepsilon_{X^2}, \varepsilon_{X^1}, \varepsilon_R \sim \mathcal{N}(0,1)$, A takes values in the space $\{a_1, \ldots, a_L\}$. In our experiments, we consider 3 possible actions ($L = 3$) and randomly draw the parameters $\beta_{a_1,1}, \ldots, \beta_{a_3,1}, \beta_{a_1,2}, \ldots, \beta_{a_3,2}$ from $\mathcal{N}(0,1)$, while the environment-specific parameters γ_e, α_e are drawn from $\mathcal{N}(0,4)$. These parameters are then fixed across all experiment runs.

D.9.2. Initial Policy

We construct an initial policy π^0 in Section 5 as follows. First, we generate a training data $D \coloneqq \{(X_i^1, X_i^2, A_i, R_i, e_i)\}_{i=1}^n$ from the uniform random policy and partition the dataset D according to the action values: $D_{a_1}, \ldots, D_{a_L}$. Then, for each action $a \in \{a_1, a_2, a_3\}$, we fit a linear regression on D_a to estimate the reward R from X^1 and X^2. Denote the resulting regressor as f_a. The initial policy is then constructed as

$$\pi^0(A = a \mid X^1, X^2) \propto \exp \frac{1}{2} f_a(X^1, X^2).$$

D.9.3. Invariant Test with True Conditional Expectation

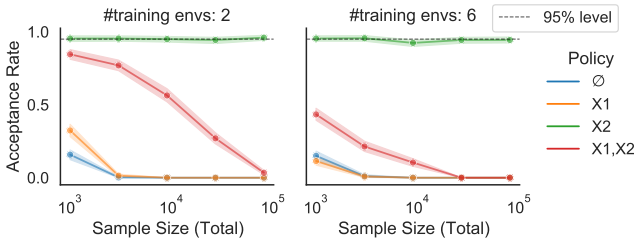


Figure D.1.: Acceptance rates for the off-policy invariance test with true conditional expectation.

D.10. Warfarin Case Study

D.10.1. Initial Policy

We generate the training data $\{(X_i, A_i, R_i, e_i)\}_{i=1}^n$, where $e_i \in \mathcal{E} = \{1, \ldots, 4\}$ under the following initial policy. We fit a linear regression to estimate the optimal warfarin dose from BMI score. Let us denote the resulting regressor by f^{BMI}. The initial policy π^0 then selects actions according to the following (unnormalized) distribution:

$$\pi^0(A = a \mid X^{\text{BMI}}) \propto \exp \frac{1}{2} \left| f^{\text{BMI}}(X^{\text{BMI}}) - m(a) \right|^{-1},$$

where, as before, $m(a)$ denotes a median value of the optimal warfarin doses within the bucket a.

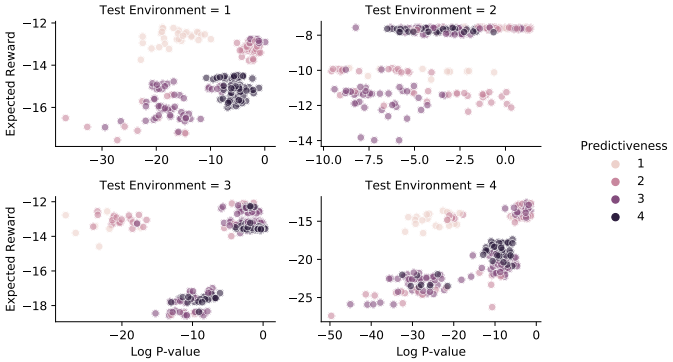


Figure D.2.: Analysis on the generalization performance and the degree of invariance. The y-axes represent the expected reward of policies with different subsets, while the x-axes represent their corresponding p-values return from the invariance test. The result shows that a policy that depends on a subset with a higher degree of invariance is more likely to generalize better to a new environment.

D.10.2. Defining Sets

The resulting defining set is {Race, VKORC1}. The following are the details of these variables (see also Consortium [2009]):

- VKORC1: Genetic information – vitamin K epoxide reductase complex, subunit 1.
- Race: Racial categories as defined by the U.S. Office of Management and Budget.

D.10.3. P-value and Generalization Analysis

In the semi-real experiment (see Section 6.4), we further analyze the generalization performance of each candidate set and its corresponding p-value returned by the invariance test. To distinguish the effects of invariance and predictiveness on the generalization performance (measured by the expected reward on a test environment), we partition the subsets into four groups depending on their performance on the training environments (1 is the least predictive and 4 is the most predictive).

Within each predictiveness group, the scatter plots in Fig. D.2 display a correlation between the p-value returned by the invariance test and the expected reward under a test environment. This result indicates that a policy that depends on a subset with a higher degree of invariance (higher p-value) tends to generalizes better to a new environment. The correlation is strongest in the test environment $e = 4$ in which we could also observe the largest performance gap between invariant and non-invariant approaches, see Fig. 5.

D.11. Example justifying Assumption 1

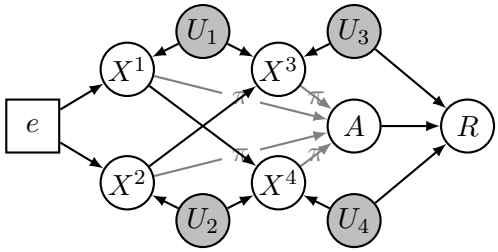


Figure D.3.: Example setting illustrating that Assumption 1 is required to derive the theoretical results in Proposition 1 and Item 2(ii).

We now discuss an example (presented in Fig. D.3) that justifies Assumption 1. In this example, the variables U_1 and U_2 influence only the observed covariates X but not the reward R. This example would lead to the following problems in Proposition 1 and Theorem 2.

First, the subsets $\{X^1, X^4\}$ and $\{X^2, X^3\}$ are both d-invariant, but no set of size 3 or more is d-invariant. By symmetry, there is no guarantee that a d-invariant set that is optimal in the training environments will also be optimal in a new test environment because e.g. $\{X^1, X^4\}$ might be optimal on the training data while $\{X^2, X^3\}$ is optimal on the test data. This then refutes the statement in Proposition 1. Assumption 1 fixes this problem as it ensures the existence of a largest d-invariant set which is a superset of all other d-invariant sets (see the proof of Lemma D.2), and rules out this example.

Second, there is no strongly non-d-invariant variable (see Definition 5) in this example and hence Assumption 2 does not guarantee the existence of a confounding removing environment. This implies that a set $\mathcal{E}$ of environments can be arbitrary, for instance, it could be a singleton $\mathcal{E} = \{e\}$. In that case, Item 2(ii) would no longer hold (but Item 2(i) remains valid). We, therefore, require Assumption 1 for proving the results of Proposition 1 and the second statement of Item 2(ii).

E. Appendix to Invariant Ancestry Search

E.1. Proofs

E.1.1. A direct Proof of Proposition 1

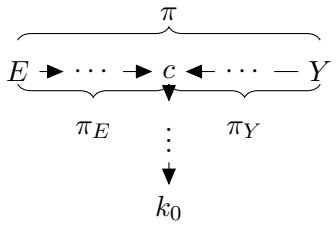
Proof. Assume that E is exogenous. If $E \in \mathrm{PA}_Y$, then there are no minimally invariant sets, and the statement holds trivially. If $E \notin \mathrm{PA}_Y$, then assume for contradiction, that an invariant set $S_0 \subsetneq S$ exists. By assumption, $|S \setminus S_0| > 1$, because otherwise S_0 would be non-invariant.

We can choose $S_1 \subseteq S$ and $k_0, k_1, \ldots, k_l \in S$ with $l \geq 1$ such that for all $i = 1, \ldots, l$: $k_i \notin \mathrm{DE}_{k_0}$ and

$$\begin{aligned} & S_0 \cup S_1 \cup \{k_0, \ldots, k_l\} = S && \in \mathcal{I} \\ \text{for } 0 \leq i < l: \quad & S_0 \cup S_1 \cup \{k_0, \ldots, k_i\} && \notin \mathcal{I} \\ & S_0 \cup S_1 && \in \mathcal{I}. \end{aligned}$$

This can be done by iteratively removing elements from $S \setminus S_0$, removing first the earliest elements in the causal order. The first invariant set reached in this process is then $S_0 \cup S_1$.

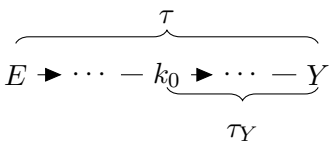
Since $S_0 \cup S_1 \cup \{k_0\}$ is non-invariant, there exists a path π between E and Y that is open given $S_0 \cup S_1 \cup \{k_0\}$ but blocked given $S_0 \cup S_1$. Since removing k_0 blocks π, k_0 must be a collider or a descendant of a collider c on π:



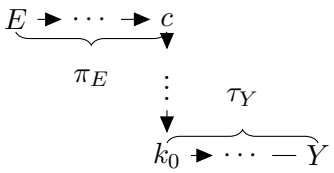
Here, $-$ represents an edge that either points left or right. Since π is open given $S_0 \cup S_1$, the two sub-paths π_E and π_Y are open given $S_0 \cup S_1$.

Additionally, since $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\} = S \setminus \{k_0\}$ is non-invariant, there exists a path τ between E and Y that is unblocked given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$ and blocked given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\} \cup \{k_0\}$. It follows that k_0 lies on τ (otherwise τ cannot be blocked by adding k_0) and k_0 has at least one outgoing edge. Assume, without loss of generality that there is an outgoing edge towards Y.

Since τ is open given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$, so is τ_Y. If there are no colliders on τ_Y, then τ_Y is also open given $S_0 \cup S_1$. But then the path the path $E \overset{\pi_E}{\cdots\!\to} c \to \cdots \to k_0 \overset{\tau_Y}{\to\!\cdots}$



is also open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.



If there are colliders on τ_Y, let m be the collider closest to k_0, meaning that $m \in \mathrm{DE}_{k_0}$. Since τ_Y is open given $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$, it means that either m or a descendant of m is in $S_0 \cup S_1 \cup \{k_1, \dots, k_l\}$. Since $\{k_1, \dots, k_l\} \cap \mathrm{DE}_{k_0} = \emptyset$, there exist $v \in (S_0 \cup S_1) \cap (\{m\} \cup \mathrm{DE}_m)$. But then $v \in \mathrm{DE}_{k_0} \cap (S_0 \cup S_1)$, meaning that π is open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.

We could assume that τ_Y had an outgoing edge from k_0 without loss of generality, because if there was instead an outgoing edge from k_0 on τ_E, the above argument would work with π_Y and τ_E instead. This concludes the proof. □

E.1.2. A direct proof of Proposition 2

Proof. If E is a parent of Y, we have $\mathcal{MI} = \emptyset$ and the statement follows trivially. Thus, assume that E is not a parent of Y. We will show that if $S \in \mathcal{I}$ is not a subset of AN_Y, then $S^* := S \cap \mathrm{AN}_Y \in \mathcal{I}$, meaning that $S \notin \mathcal{MI}$.

Assume for contradiction that there is a path p between E and Y that is open given S^*. Since $S \in \mathcal{I}$, p is blocked given S. Then there exists a non-collider Z on p that is in $S \setminus \mathrm{AN}_Y$. We now argue that all nodes on p are ancestors of Y, yielding a contradiction.

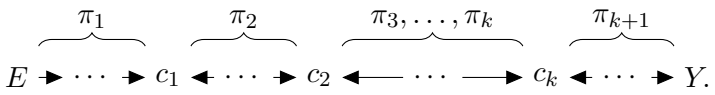
First, assume that there are no colliders on p. If E is exogenous, then p is directed from E to Y. (If E is an ancestor of Y, any node on p is either an ancestor of Y or E, and thus Y.) Second, assume that there are colliders on p. Since p is open given the smaller set $S^* \subsetneq S$, all colliders on p are in S^* or have a descendant in S^*; therefore all colliders are ancestors of Y. If E is exogenous, any node on p is either an ancestor of Y or of a collider on p. (If E is an ancestor of Y, any node on p is either an ancestor of Y, of a collider on p or of E, and thus also Y.) This completes the proof of Proposition 2. □

E.1.3. Proof of Proposition 3

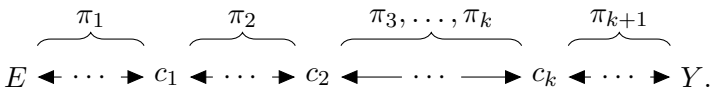
Proof. First, we show that $S_{\mathrm{IAS}} \in \mathcal{I}$. If S_{IAS} is the union of a single minimally invariant set, it trivially holds that $S_{\mathrm{IAS}} \in \mathcal{I}$. Now assume that S_{IAS} is the union of at least two

minimally invariant sets, $S_{\text{IAS}} = S_1 \cup \ldots \cup S_n$, $n \geq 2$, and assume for a contradiction that there exists a path π between E and Y that is unblocked given S_{IAS}.

Since π is blocked by a strict subset of S_{IAS}, it follows that π has at least one collider; further every collider of π is either in S_{IAS} or has a descendant in S_{IAS}, and hence every collider of π is an ancestor of Y, by Proposition 2. If E is exogenous, π has the following shape



(If E is not exogenous but $E \in \text{AN}_Y$, then π takes either the form displayed above or the shape displayed below. However, no matter which of the shapes π takes, the proof proceeds the same.)



The paths $\pi_1, \ldots, \pi_{k+1}$, $k \geq 1$, do not have any colliders and are unblocked given S_{IAS}. In particular, $\pi_1, \ldots, \pi_{k+1}$ are unblocked given S_1.

The path π_{k+1} must have a final edge pointing to Y, because otherwise it would be a directed path from Y to c_k, which contradicts acyclicity since c_k is an ancestor of Y.

As c_1 is an ancestor of Y, there exists a directed path, say ρ_1, from c_1 to Y. Since π_1 is open given S_1 and since S_1 is invariant, it follows that ρ_1 must be blocked by S_1 (otherwise the path $E \overset{\pi_1}{\to} c_1 \overset{\rho_1}{\to} Y$ would be open). For this reason, S_1 contains a descendant of the collider c_1.

Similarly, if ρ_2 is a directed path from c_2 to Y, then S_1 blocks ρ_2, because otherwise the path $E \overset{\pi_1}{\to} c_1 \leftarrow \overset{\pi_2}{\cdots} \to c_2 \overset{\rho_2}{\to} Y$ would be open. Again, for this reason, S_1 contains a descendant of c_2.

Iterating this argument, it follows that S_1 contains a descendant of every collider on π, and since $\pi_1, \ldots, \pi_{k+1}$ are unblocked by S_1, π is open given S_1. This contradicts invariance of S_1 and proves that $S_{\text{IAS}} \in \mathcal{I}$.

We now show that $S_{\text{ICP}} \subseteq S_{\text{IAS}}$ with equality if and only if $S_{\text{ICP}} \in \mathcal{I}$. First, $S_{\text{ICP}} \subseteq S_{\text{IAS}}$ because S_{IAS} is a union of the minimally invariant sets, and S_{ICP} is the intersection over all invariant sets. We now show the equivalence statement.

Assume first that $S_{\text{ICP}} \in \mathcal{I}$. As S_{ICP} is the intersection of all invariant sets, $S_{\text{ICP}} \in \mathcal{I}$ implies that there exists exactly one invariant set, that is contained in all other invariant sets. By definition, this means that there is only one minimally invariant set, and that this set is exactly S_{ICP}. Thus, $S_{\text{IAS}} = S_{\text{ICP}}$.

Conversely assume that $S_{\text{ICP}} \notin \mathcal{I}$. By construction, S_{ICP} is contained in any invariant set, in particular in the minimally invariant sets. However, since S_{ICP} is not invariant itself, this containment is strict, and it follows that $S_{\text{ICP}} \subsetneq S_{\text{IAS}}$.

□

E.1.4. Proof of Proposition 4

Proof. First we show $\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq S_{\mathrm{ICP}}$. If $j \in \mathrm{PA}_Y \cap \mathrm{CH}_E$, any invariant set contains j, because otherwise the path $E \to j \to Y$ is open. Similarly, if $j \in \mathrm{PA}_Y \cap \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)$, any invariant set contains j (there exists a node j' such that $E \to j' \to \cdots \to Y$ and $E \to j' \leftarrow j \to Y$, and any invariant set S must contain j' or one of its descendants; thus, it must also contain j to ensure that the path $E \to j' \leftarrow j \to Y$ is blocked by S.) It follows that for all invariant S,

$$\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq S,$$

such that

$$\mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)) \subseteq \bigcap_{S \text{ invariant}} S.$$

To show $S_{\mathrm{ICP}} \subseteq \mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E))$, take any $j \notin \mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E))$. We argue, that an invariant set $\bar{S}$ not containing j exists, such that $j \notin S_{\mathrm{ICP}} = \bigcap_{S \text{ invariant}} S$. If $j \notin \mathrm{PA}_Y$, let $\bar{S} = \mathrm{PA}_Y$, which is invariant. If $j \in \mathrm{PA}_Y$, define

$$\bar{S} = (\mathrm{PA}_Y \setminus \{j\}) \cup \mathrm{PA}_j \cup (\mathrm{CH}_j \cap \mathrm{AN}_Y) \cup \mathrm{PA}(\mathrm{CH}_j \cap \mathrm{AN}_Y).$$

Because $j \notin \mathrm{CH}_E$ and $j \notin \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E)$, we have $E \notin \bar{S}$. Also observe that $\bar{S} \subseteq \mathrm{AN}_Y$. We show that any path between E and Y is blocked by $\bar{S}$, by considering all possible paths:

$\cdots \mathbf{j'} \to \mathbf{Y}$ **for** $\mathbf{j' \neq j}$**:** Blocked because $j' \in \mathrm{PA}_Y \setminus \{j\}$.

$\cdots \mathbf{v} \to \mathbf{j} \to \mathbf{Y}$**:** Blocked because $v \in \mathrm{PA}_j \subseteq \bar{S}$ and $E \notin \mathrm{PA}_j$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \mathrm{PA}_j(\mathrm{CH}_j \cap \mathrm{AN}_Y)$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $\bar{S} \subseteq \mathrm{AN}_Y$, and since $c \notin \mathrm{AN}_Y$, $\bar{S} \cap \mathrm{DE}_c = \emptyset$ and the path is blocked given $\bar{S}$ because of the collider c.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \mathrm{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \mathrm{AN}_c$ and $c \in \mathrm{AN}_Y$, so $v \in \mathrm{CH}_j \cap \mathrm{AN}_Y \subseteq \bar{S}$.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$**:** Same reason as for the case '$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ and $\mathbf{c} \notin \mathrm{AN}_{\mathbf{Y}}$'.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{Y}$ Since $\bar{S} \subseteq \mathrm{AN}_Y$, we must have $\bar{S} \cap \mathrm{DE}_c = \emptyset$ (otherwise this would create a directed cycle from $Y \to \cdots \to Y$). Hence the path is blocked given $\bar{S}$ because of the collider c.

Since there are no open paths from E to Y given $\bar{S}$, $\bar{S}$ is invariant, and $S_{\mathrm{ICP}} \subseteq \bar{S}$. Since $j \notin \bar{S}$, it follows that $j \notin S_{\mathrm{ICP}}$. This concludes the proof. □

E.1.5. Proof of Theorem 1

Proof. Consider first the case where all marginal tests have pointwise asymptotic power and pointwise asymptotic level.

Pointwise asymptotic level: Let $\mathbb{P}_0 \in H_{0,S}^{\mathcal{MI}}$. By the assumption of pointwise asymptotic level, there exists a non-negative sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \varepsilon_n = 0$ and $\mathbb{P}_0(\phi_n(S) = 1) \leq \alpha + \varepsilon_n$. Then

$$\begin{aligned}
\mathbb{P}_0(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S \setminus \{j\}) = 0)\right) \\
&\leq \mathbb{P}_0(\phi_n(S) = 1) + \sum_{j\in S} \mathbb{P}_0(\phi_n(S \setminus \{j\}) = 0) \\
&\leq \alpha + \varepsilon_n + \sum_{j\in S} \mathbb{P}_0(\phi_n(S \setminus \{j\}) = 0) \\
&\to \alpha + 0 \qquad \text{as } n \to \infty \\
&= \alpha.
\end{aligned}$$

The convergence step follows from

$$H_{0,S}^{\mathcal{MI}} = H_{0,S}^{\mathcal{I}} \cap \bigcap_{j\in S} H_{A,S\setminus\{j\}}^{\mathcal{I}}$$

and from the assumption of pointwise asymptotic level and power. As $\mathbb{P}_0 \in H_{0,S}^{\mathcal{MI}}$ was arbitrary, this shows that $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level.

Pointwise asymptotic power: To show that the decision rule has pointwise asymptotic power, consider any $\mathbb{P}_A \in H_{A,S}^{\mathcal{MI}}$. We have that

$$H_{A,S}^{\mathcal{MI}} = H_{A,S}^{\mathcal{I}} \cup \left(H_{0,S}^{\mathcal{I}} \cap \bigcup_{j\in S} H_{0,S\setminus\{j\}}^{\mathcal{I}}\right). \tag{E.1}$$

As the two sets $H_{A,S}^{\mathcal{I}}$ and

$$H_{0,S}^{\mathcal{I}} \cap \bigcup_{j\in S} H_{0,S\setminus\{j\}}^{\mathcal{I}}$$

are disjoint, we can consider them one at a time. Consider first the case $\mathbb{P}_A \in H_{A,S}^{\mathcal{I}}$. This means that S is not invariant and thus

$$\begin{aligned}
\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_A\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S \setminus \{j\}, \alpha) = 0)\right) \\
&\geq \mathbb{P}_A(\phi_n(S) = 1) \\
&\to 1 \qquad \text{as } n \to \infty
\end{aligned}$$

by the assumption of pointwise asymptotic power.

Next, assume that there exists $j' \in S$ such that $\mathbb{P}_A \in (H^{\mathcal{I}}_{0,S} \cap H^{\mathcal{I}}_{0,S\setminus\{j'\}})$. Then,

$$\begin{aligned}\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S\setminus\{j\}) = 0)\right)\\ &\geq \mathbb{P}_A(\phi_n(S\setminus\{j'\}) = 0)\\ &\geq 1 - \alpha - \varepsilon_n\\ &\to 1-\alpha \qquad \text{as } n\to\infty.\end{aligned}$$

Thus, for arbitrary $\mathbb{P}_A \in H^{\mathcal{MI}}_{A,S}$ we have shown that $\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) \geq 1-\alpha$ in the limit. This shows that $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic power of at least $1-\alpha$. This concludes the argument for pointwise asymptotic power.

Next, consider the case that the marginal tests have uniform asymptotic power and uniform asymptotic level. The calculations for showing that $\phi_n^{\mathcal{MI}}$ has uniform asymptotic level and uniform asymptotic power of at least $1-\alpha$ are almost identical to the pointwise calculations.

Uniform asymptotic level: By the assumption of uniform asymptotic level, there exists a non-negative sequence ε_n such that $\lim_{n\to\infty}\varepsilon_n = 0$ and $\sup_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}}\mathbb{P}(\phi_n(S) = 1) \leq \alpha + \varepsilon_n$. Then,

$$\begin{aligned}\sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) &= \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\mathbb{P}\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S\setminus\{j\}) = 0)\right)\\ &\leq \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\left(\mathbb{P}(\phi_n(S) = 1) + \sum_{j\in S}\mathbb{P}(\phi_n(S\setminus\{j\}) = 0)\right)\\ &\leq \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\mathbb{P}(\phi_n(S) = 1) + \sum_{j\in S}\sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\mathbb{P}(\phi_n(S\setminus\{j\}) = 0)\\ &\leq \alpha + \varepsilon_n + \sum_{j\in S}\left(1 - \inf_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}}\mathbb{P}(\phi_n(S\setminus\{j\}) = 1)\right)\\ &\to \alpha + 0 + \sum_{j\in S}(1-1) \qquad \text{as } n\to\infty\\ &= \alpha.\end{aligned}$$

Uniform asymptotic power: From (E.1), it follows that

$$\inf_{\mathbb{P}\in H^{\mathcal{MI}}_{A,S}}\mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) = \min\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}}\mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1), \inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap\bigcup_{j\in S}H^{\mathcal{I}}_{0,S\setminus\{j\}}}\mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1)\right\}.$$

We consider the two inner terms in the above separately. First,

$$\begin{aligned}\inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1) &= \inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\\ &\geq \inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}(\phi_n(S)=1)\\ &\to 1 \qquad \text{as } n\to\infty.\end{aligned}$$

Next,

$$\begin{aligned}&\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap\bigcup_{j\in S}H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S)=1)\\ &= \inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap\bigcup_{j\in S}H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\\ &= \min_{j\in S}\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}\left((\phi_n(S)=1)\cup\bigcup_{j\in S}(\phi_n(S\setminus\{j\})=0)\right)\right\}\\ &\geq \min_{j\in S}\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n(S\setminus\{j\})=0)\right\}\\ &= \min_{j\in S}\left\{1-\sup_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}\cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n(S\setminus\{j\})=1)\right\}\\ &\geq 1-\alpha-\varepsilon_n\\ &\to 1-\alpha \qquad \text{as } n\to\infty.\end{aligned}$$

This shows that $\phi_n^{\mathcal{MI}}$ has uniform asymptotic power of at least $1-\alpha$, which completes the proof. □

E.1.6. Proof of Theorem 2

Proof. We have that

$$\lim_{n\to\infty}\mathbb{P}(\hat{S}_{\text{IAS}}\subseteq \text{AN}_Y)\geq\lim_{n\to\infty}\mathbb{P}(\hat{S}_{\text{IAS}}=S_{\text{IAS}})$$

as $S_{\text{IAS}}\subseteq \text{AN}_Y$ by Proposition 3. Furthermore, we have

$$\mathbb{P}(\hat{S}_{\text{IAS}}=S_{\text{IAS}})\geq\mathbb{P}(\widehat{\mathcal{MI}}=\mathcal{MI}).$$

Let $A \coloneqq \{S \mid S \notin \mathcal{I}\} \setminus \{S \mid \exists S' \subsetneq S \text{ s.t. } S' \in \mathcal{MI}\}$ be those non-invariant sets that do not contain a minimally invariant set and observe that

$$(\widehat{\mathcal{MI}} = \mathcal{MI}) \supseteq \bigcap_{S \in \mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S \in A} (\phi_n(S, \alpha C^{-1}) = 1). \tag{E.2}$$

To see why this is true, note that to correctly recover $\mathcal{MI}$, we need to 1) accept the hypothesis of minimal invariance for all minimally invariant sets and 2) reject the hypothesis of invariance for all non-invariant sets that are not supersets of a minimally invariant set (any superset of a set for which the hypothesis of minimal invariance is not rejected is removed in the computation of $\widehat{\mathcal{MI}}$). Then,

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{MI}} = \mathcal{MI}) &\geq \mathbb{P}\left(\bigcap_{S \in \mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S \in A} (\phi_n(S, \alpha C^{-1}) = 1) \right) \\
&\geq 1 - \mathbb{P}\left(\bigcup_{S \in \mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 1) \right) - \sum_{S \in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S \in \mathcal{MI}} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 1) - \sum_{S \in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S \in \mathcal{MI}} (\alpha C^{-1} + \varepsilon_{n,S}) - \sum_{S \in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - |\mathcal{MI}| \alpha C^{-1} + \sum_{S \in \mathcal{MI}} \varepsilon_{n,S} - \sum_{S \in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \alpha + \sum_{S \in \mathcal{MI}} \varepsilon_{n,S} - \sum_{S \in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\to 1 - \alpha \quad \text{as } n \to \infty,
\end{aligned}$$

where $(\varepsilon_{n,S})_{n \in \mathbb{N}, S \in \mathcal{MI}}$ are non-negative sequences that converge to zero and the last step follows from the assumption of asymptotic power. The sequences $(\varepsilon_{n,S})_{n \in \mathbb{N}, S \in \mathcal{MI}}$ exist by the assumption of asymptotic level. □

E.1.7. Proof of Proposition 5

Proof. We prove the statements one by one.

(i) Since S^m_{IAS} is the union over some of the minimally invariant sets, $S^m_{\text{IAS}} \subseteq S_{\text{IAS}}$. Then the statement follows from Proposition 2.

(ii) If $m \geq m_{\max}$, all $S \in \mathcal{MI}$ satisfy the requirement $|S| \leq m$.

(iii) If $m \geq m_{\min}$, then S^m_{IAS} contains at least one minimally invariant set. The statement then follows from the first part of the proof of Proposition 3 given in Appendix E.1.3.

(iv) S^m_{IAS} contains at least one minimally invariant set and, by (iii), it is itself invariant. Thus, if $S_{\text{ICP}} \notin \mathcal{I}$, then $S_{\text{ICP}} \subsetneq S^m_{\text{IAS}}$. If $S_{\text{ICP}} \in \mathcal{I}$, then there exists only one minimally invariant set, which is S_{ICP} (see proof of Proposition 3), and we have $S_{\text{ICP}} = S^m_{\text{IAS}}$. This concludes the proof. □

E.1.8. Proof of Theorem 3

Proof. The proof is identical to the proof of Theorem 2, when changing the correction factor 2^{-d} to $C(m)^{-1}$, adding superscript m's to the quantities $\widehat{\mathcal{MI}}$, $\hat{S}_{\text{IAS}}$ and S_{IAS}, and adding the condition $|S| \leq m$ to all unions, intersections and sums. □

E.1.9. Proof of Proposition 6

By Proposition 2, we have $S_{\text{IAS}} \subseteq \text{AN}_Y$, and since $S_{\text{IAS},O} \subseteq S_{\text{IAS}}$, the claim follows immediately.

E.2. Oracle Algorithms for Learning S_{IAS}

In this section, we review some of the existing literature on minimal d-separators, which can be exploited to give an algorithmic approach for finding S_{IAS} from a DAG. We first introduce the concept of M-minimal separation with respect to a constraining set I.

Definition E.1 (van der Zander et al. [2019], Section 2.2)**.** Let $I \subseteq [d]$, $K \subseteq [d]$, and $S \subseteq [d]$. We say that S is a *K-minimal separator* of E and Y with respect to a constraining set I if all of the following are true:

(i) $I \subseteq S$.

(ii) $S \in \mathcal{I}$.

(iii) There does not exists $S' \in \mathcal{I}$ such that $K \subseteq S' \subsetneq S$.

We denote by $M_{K,I}$ the set of all K-minimal separating sets with respect to constraining set I.

(In this work, $S \in \mathcal{I}$ means $E \perp\!\!\!\perp Y \mid S$, but it can stand for other separation statements, too.) The definition of a K-minimal separator coincides with the definition of a minimally invariant set if both K and the constraining set I are equal to the empty set. An $\emptyset$-minimal separator with respect to constraining set I is called a *strongly-minimal separator with respect to constraining set I*.

We can now represent (2) using this notation. $M_{\emptyset,\emptyset}$ contains the minimally invariant sets and thus

$$S_{\text{IAS}} := \bigcup_{S \in M_{\emptyset,\emptyset}} S.$$

Listing the set $M_{I,I}$ of all I-minimal separators with respect to the constraining set I (for any I) can be done in polynomial delay time $\mathcal{O}(d^3)$ [van der Zander et al., 2019, Takata,

2010], where delay here means that finding the next element of $M_{I,I}$ (or announcing that there is no further element) has cubic complexity. This is the algorithm we exploit, as described in the main part of the paper.

Furthermore, we have

$$i \in S_{\text{IAS}} \quad \Leftrightarrow \quad M_{\emptyset,\{i\}} \neq \emptyset.$$

This is because $i \in S_{\text{IAS}}$ if and only if there is a minimally invariant set that contains i, which is the case if and only if there exist a strongly minimal separating set with respect to constraining set $\{i\}$. Thus, we can construct S_{IAS} by checking, for each i, whether there is an element in $M_{\emptyset,\{i\}}$. Finding a strongly-minimal separator with respect to constraining set I, i.e., finding an element in $M_{\emptyset,I}$, is NP-hard if the set I is allowed to grow [van der Zander et al., 2019]. To the best of our knowledge, however, it is unknown whether finding an element in $M_{\emptyset,\{i\}}$, for a singleton $\{i\}$ is NP-hard.

E.3. The Maximum Number of Minimally Invariant Sets

If one does not have a priori knowledge about the graph of the system being analyzed, one can still apply Theorem 2 with a correction factor 2^d, as this ensures (with high probability) that no minimally invariant sets are falsely rejected. However, we know that the correction factor is strictly conservative, as there cannot exist 2^d minimally invariant sets in a graph. Thus, correcting for 2^d tests, controls the familywise error rate (FWER) among minimally invariant sets, but increases the risk of falsely accepting a non-invariant set relatively more than what is necessary to control the FWER. Here, we discuss the maximum number of minimally invariant sets that can exist in a graph with d predictor nodes and how a priori knowledge about the sparsity of the graph and the number of interventions can be leveraged to estimate a less strict correction that still controls the FWER.

As minimally invariant sets only contain ancestors of Y (see Proposition 2), we only need to consider graphs where Y comes last in a causal ordering. Since d-separation is equivalent to undirected separation in the moralized ancestral graph [Lauritzen, 1996], finding the largest number of minimally invariant sets is equivalent to finding the maximum number of minimal separators in an undirected graph with $d+2$ nodes. It is an open question how many minimal separators exists in a graph with $d+2$ nodes, but it is known that a lower bound for the maximum number of minimal separators is in $\Omega(3^{d/3})$ [Gaspers and Mackenzie, 2015]. We therefore propose using a correction factor of $C = 3^{\lceil d/3 \rceil}$ when estimating the set $\hat{S}_{\text{IAS}}$ from Theorem 2 if one does not have a priori knowledge of the number of minimally invariant sets in the DAG of the SCM being analyzed. This is a heuristic choice and is not conservative for all graphs.

Theorem 2 assumes asymptotic power of the invariance test, but as we can only have a finite amount of data, we will usually not have full power against all non-invariant sets that are not supersets of a minimally invariant set. Therefore, choosing a correction factor that is potentially too low represents a trade-off between error types: if we correct too little, we stand the risk of falsely rejecting a minimally invariant set but not rejecting a superset of it, whereas when correcting too harshly, there is a risk of failing to reject

non-invariant sets due to a lack of power.

If one has a priori knowledge of the sparsity or the number of interventions, these can be leveraged to estimate the maximum number of minimally invariant sets using simulation, by the following procedure:

1. For $b = 1, \ldots, B$:
   a) Sample a DAG with d predictor nodes, $N_{\text{interventions}} \sim \mathbb{P}_N$ interventions and $p \sim \mathbb{P}_p$ probability of an edge being present in the graph over (X, Y), such that Y is last in a causal ordering. The measures $\mathbb{P}_N$ and $\mathbb{P}_p$ are distributions representing a priori knowledge. For instance, in a controlled experiment, the researcher may have chosen the number N_0 of interventions. Then, $\mathbb{P}_N$ is a degenerate distribution with $\mathbb{P}_N(N_0) = 1$.
   b) Compute the set of all minimally invariant sets, e.g., using the `adjustmentSets` algorithm from `dagitty` [Textor et al., 2016].
   c) Return the number of minimally invariant sets.
2. Return the largest number of minimally sets found in the B repetitions above.

Instead of performing B steps, one can continually update the largest number of minimally invariant sets found so far and end the procedure if the maximum has not updated in a predetermined number of steps, for example.

E.4. A Finite Sample Algorithm for Computing $\hat{S}_{\text{IAS}}$

In this section, we provide an algorithm for computing the sets $\hat{S}_{\text{IAS}}$ and $\hat{S}^m_{\text{IAS}}$ presented in Theorems 2 and 3. The algorithm finds minimally invariant sets by searching for invariant sets among sets of increasing size, starting from the empty set. This is done, because the first (correctly) accepted invariant is a minimally invariant set. Furthermore, any set that is a superset of an accepted invariant set, does not need to be tested (as this set cannot be minimal). Tests for invariance can be computationally expensive if one has large amounts of data. Therefore, skipping unnecessary tests offers a significant speedup. In the extreme case, where all singletons are found to be invariant, the algorithm completes in $d + 1$ steps, compared to $\sum_{i=0}^{m} \binom{d}{i}$ steps (2^d if $m = d$). This is implemented in lines 8-10 of Algorithm E.1.

E.5. Additional Experiment Details

E.5.1. Simulation Details for Section 6.1

We sample graphs that satisfy Assumption 1 with the additional requirement that $Y \in \text{DE}_Y$ by the following procedure:

1. Sample a DAG $\mathcal{G}$ for the graph of (X, Y) with $d+1$ nodes, for $d \in \{4, 6, \ldots, 20\} \cup \{100, 1{,}000\}$, and choose Y to be a node (chosen uniformly at random) that is not a root node.

Algorithm E.1 An algorithm for computing $\hat{S}_{\text{IAS}}$ from data

Input: A decision rule ϕ_n for invariance, significance thresholds α_0, α, max size of sets to test m (potentially $m = d$) and data
Output: The set $\hat{S}_{\text{IAS}}$

Initialize $\widehat{\mathcal{MI}}$ as an empty list.
$PS \leftarrow \{S \subseteq [d] \mid |S| \leq m\}$
if $\phi_n(\emptyset, \alpha_0) = 0$ **then**
    End the procedure and return $\hat{S}_{\text{IAS}} = \emptyset$
Sort PS in increasing order according the set sizes
for $S \in PS$ **do**
    if $S \supsetneq S'$ for any $S' \in \widehat{\mathcal{MI}}$ **then**
        Skip the test of S and go to next iteration of the loop
    else
        Add S to $\widehat{\mathcal{MI}}$ if $\phi_n(S, \alpha) = 0$, else continue
    if The union of $\widehat{\mathcal{MI}}$ contains all nodes **then**
        Break the loop
Return $\hat{S}_{\text{IAS}}$ as the union of all sets in $\widehat{\mathcal{MI}}$

2. Add a root node E to $\mathcal{G}$ with $N_{\text{interventions}}$ children that are not Y. When $d \leq 20$, $N_{\text{interventions}} \in \{1, \dots, d\}$ and when $d \geq 100$, $N_{\text{interventions}} \in \{1, \dots, 0.1 \times d\}$ (i.e., we consider interventions on up to ten percent of the predictor nodes).

3. Repeat the first two steps if $Y \notin \text{DE}_E$.

E.5.2. Simulation Details for Section 6.2

We simulate data for the experiment in Section 6.2 (and the additional plots in Appendix E.5.4) by the following procedure:

1. Sample data from a single graph by the following procedure:
    a) Sample a random graph $\mathcal{G}$ of size $d + 1$ and sample Y (chosen uniformly at random) as any node that is not a root node in this graph.
    b) Sample coefficients, $\beta_{i \to j}$, for all edges $(i \to j)$ in $\mathcal{G}$ from $U((-2, 0.5) \cup (0.5, 2))$ independently.
    c) Add a node E with no incoming edges and $N_{\text{interventions}}$ children, none of which are Y. When $d = 6$, we set $N_{\text{interventions}} = 1$ and when $d = 100$, we sample $N_{\text{interventions}}$ uniformly from $\{1, \dots, 10\}$.
    d) If Y is not a descendant of E, repeat steps (a), (b) and (c) until a graph where $Y \in \text{DE}_E$ is obtained.
    e) For $n \in \{10^2, 10^3, 10^4, 10^5\}$:
        i. Draw 50 datasets of size n from an SCM with graph $\mathcal{G}$ and coefficients $\beta_{i \to j}$ and with i.i.d. $N(0, 1)$ noise innovations. The environment variable,

E, is sampled independently from a Bernoulli distribution with probability parameter $p = 0.5$, corresponding to (roughly) half the data being observational and half the data interventional. The data are generated by looping through a causal ordering of (X, Y), starting at the bottom, and standardizing a node by its own empirical standard deviation before generating children of that node; that is, a node X_j is first generated from PA_j and then standardized before generating any node in CH_j. If X_j is intervened on, we standardize it prior to the intervention.

ii. For each sampled dataset, apply IAS and ICP. Record the Jaccard similarities between IAS and AN_Y and between ICP and AN_Y, and record whether or not is was a subset of AN_Y and whether it was empty.

iii. Estimate the quantity plotted (average Jaccard similarity in Fig. 4 or probability of $\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$ or $\hat{S}_{\mathrm{IAS}} = \emptyset$ in Fig. E.2) from the 50 simulated datasets.

f) Return the estimated quantities from the previous step.

2. Repeat the above 100 times and save the results in a data-frame.

E.5.3. Analysis of the Choice of C in Section 6.2

We have repeated the simulation with $d = 6$ from Section 6.2 but with a correction factor of $C = 2^6$, as suggested by Theorem 2 instead of the heuristic correction factor of $C = 9$ suggested in Appendix E.3. Fig. E.1 shows the results. We see that the results are almost identical to those presented in Fig. 4. Thus, in the scenario considered here, there is no change in the performance of $\hat{S}_{\mathrm{IAS}}$ (as measured by Jaccard similarity) between using a correction factor of $C = 2^6$ and a correction factor of $C = 3^{\lceil 6/3 \rceil} = 9$. In larger graphs, it is likely that there is a more pronounced difference. E.g., at $d = 10$, the strictly conservative correction factor suggested by Theorem 2 is $2^{10} = 1024$, whereas the correction factor suggested in Appendix E.3 is only $3^{\lceil 10/3 \rceil} = 3^4 = 81$, and at $d = 20$ the two are $2^{20} = 1{,}048{,}576$ and $3^{\lceil 20/3 \rceil} = 3^7 = 2187$.

E.5.4. Analysis of the Choice of α_0 in Section 6.2

Here, we investigate the quantities $\mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$, $\mathbb{P}(\hat{S}^1_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$, $\mathbb{P}(\hat{S}_{\mathrm{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}^1_{\mathrm{IAS}} = \emptyset)$ using the same simulation setup as described in Section 6.2. Furthermore, we also ran the simulations for values $\alpha_0 = \alpha$ (testing all hypotheses at the same level), $\alpha_0 = 10^{-6}$ (conservative, see Remark 1) as in Section 6.2 and $\alpha_0 = 10^{-12}$ (very conservative). The results for $\alpha = 10^{-6}$ (shown in Fig. E.2) were recorded in the same simulations that produced the output for Fig. 4. For $\alpha_0 \in \{\alpha, 10^{-12}\}$ (shown in Fig. E.3 and Fig. E.4, respectively) we only simulated up to 10,000 observations, to keep computation time low.

Generally, we find that the probability of IAS being a subset of the ancestors seems to generally hold well and even more so with large sample sizes. (see Figs. E.2 to E.4), in line with Theorem 2. When given 100,000 observations, the probability of IAS being

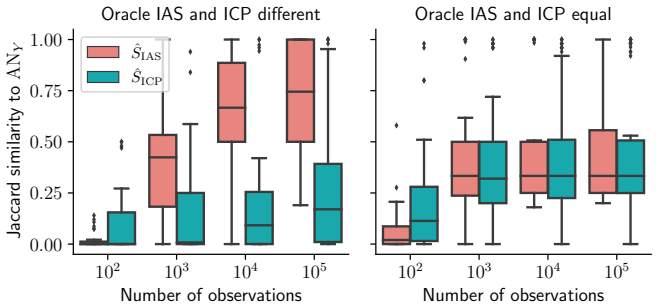


Figure E.1.: The same figure as in Fig. 4, but with a correction factor of $C = 2^6 = 64$ instead of $C = 3^{\lceil 6/3 \rceil} = 9$. Only $d = 6$ shown here, as the correction factor for $d = 100$ is unchanged. Here, the guarantees of Theorem 2 are not violated by a potentially too small correction factor, and the results are near identical to those given in Fig. 4 using a milder correction factor.

a subset of ancestors is roughly equal to one for almost all SCMs, although there are a few SCMs, where IAS is never a subset of the ancestors (see Fig. E.2). For $\alpha_0 = 10^{-6}$, the median probability of IAS containing only ancestors is one in all cases, except for $d = 100$ with 1,000 observations – here, the median probability is 87%.

In general, varying α_0 has the effect hypothesized in Remark 1: lowering α_0 increases the probability that IAS contains only ancestors, but at the cost of increasing the probability that it is empty (see Figs. E.2 to E.4). For instance, the median probability of IAS being a subset of ancestors when $\alpha_0 = 10^{-12}$ is one for all sample sizes, but the output is always empty when there are 100 observations and empty roughly half the time even at 1,000 observations when $d = 100$ (see Fig. E.4). In contrast, not testing the empty set at a reduced level, means that the output of IAS is rarely empty, but the probability of IAS containing only ancestors decreases. Still, even with $\alpha_0 = \alpha$, the median probability of IAS containing only ancestors was never lower than 80% (see Fig. E.3). Thus, choosing α_0 means choosing a trade-off between finding more ancestor-candidates, versus more of them being false positives.

E.5.5. Analysis of the strength of inverventions in Section 6.2

Here, we repeat the $d = 6$ simulations from Section 6.2 with a reduced strength of the environment to investigate the performance of IAS under weaker interventions. We sample from the same SCMs as sampled in Section 6.2, but reduce the strength of the interventions to be 0.5 instead of 1. That is, the observational distributions are the same as in Section 6.2, but interventions to a node X_j are here half as strong as in Section 6.2.

The Jaccard similarity between $\hat{S}_{\mathrm{IAS}}$ and AN_Y is generally lower than what we found in Fig. 4 (see Fig. E.5). This is likely due to having lower power to detect non-invariance, which has two implications. First, lower power means that we may fail to reject the empty set, meaning that we output nothing. Then, the Jaccard similarity between $\hat{S}_{\mathrm{IAS}}$ and AN_Y is zero. Second, it may be that we correctly reject the empty set, but fail to

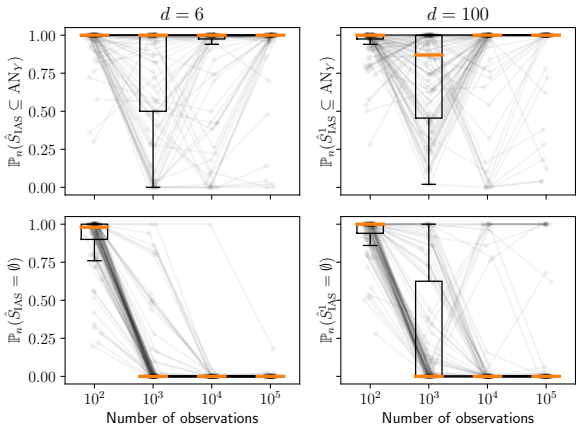


Figure E.2.: The empirical probabilities of recovering a subset of AN_Y (top row) and recovering an empty set (bottom row), when testing the empty set for invariance at level $\alpha_0 = 10^{-6}$. Generally, our methods seem to hold level well, especially when sample sizes are large. When the sample size is small, the output is often the empty set. When $d = 6$, we estimate $\hat{S}_{\text{IAS}}$ (left column) and when $d = 100$, we estimate $\hat{S}^1_{\text{IAS}}$ (right column). The results here are from the simulations that also produced Fig. 4. Medians are displayed as orange lines through each boxplot. Each point represents the probability that the output set is ancestral (resp. empty) for a randomly selected SCM, as estimated by repeatedly sampling data from the same SCM for every $n \in \{10^2, 10^3, 10^4, 10^5\}$. Observations from the same SCM are connected by a line. Each figure contains data from 100 randomly drawn SCMs. Points have been perturbed slightly along the x-axis to improve readability.

reject another non-invariant set which is not an ancestor of Y which is then potentially included in the output. Then, the $\hat{S}_{\text{IAS}}$ and AN_Y is lower, because we increase the number of false findings.

We find that the probability that $\hat{S}_{\text{IAS}}$ is a subset of ancestors is generally unchanged for the lower intervention strength, but the probability of $\hat{S}_{\text{IAS}}$ generally increases for small sample sizes (see Table E.1). This indicates that IAS does not make more mistakes under the weaker interventions, but it is more often uninformative. We see also that in both settings, $\hat{S}_{\text{IAS}}$ is empty more often than $\hat{S}_{\text{ICP}}$ for low sample sizes, but less often for larger samples (see Table E.1). This is likely because IAS tests the empty set at a much lower level than ICP does (10^{-6} compared to 0.05). Thus, IAS requires more power to find anything, but once it has sufficient power, it finds more than ICP (see also Fig. E.5). The median probability of ICP returning a subset of the ancestors was always at least 95% (not shown).

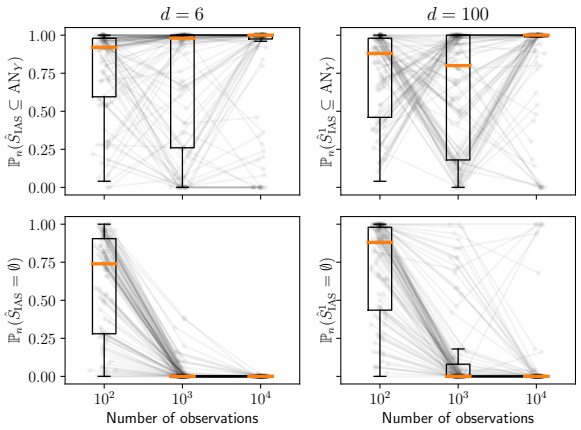


Figure E.3.: The same figure as Fig. E.2, but with $\alpha_0 = \alpha = 0.05$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at the non-conservative level $\alpha_0 = \alpha$ means that the empty set is output less often for small sample sizes, but decreases the probability that the output is a subset of ancestors. Thus, we find more ancestor-candidates, but make more mistakes when $\alpha_0 = \alpha$. However, the median probability of the output being a subset of ancestors is at least 80% in all configurations.

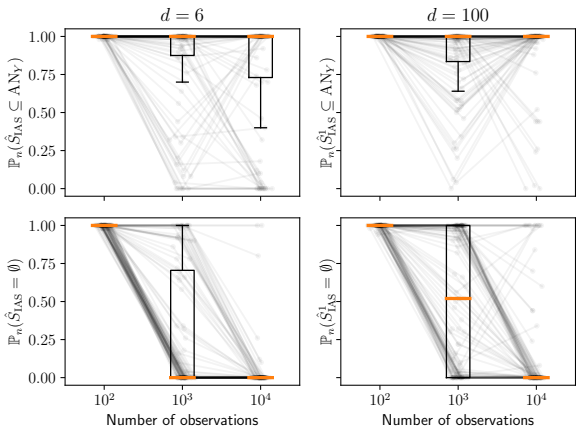


Figure E.4.: The same figure as Fig. E.2, but with $\alpha_0 = 10^{-12}$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at at very conservative level $\alpha_0 = 10^{-12}$ means that the empty set is output more often (for one hundred observations, we only find the empty set), but increases the probability that the output is a subset of ancestors. Thus, testing at a very conservative level $\alpha_0 = 10^{-12}$ means that we do not make many mistakes, but the output is often non-informative.

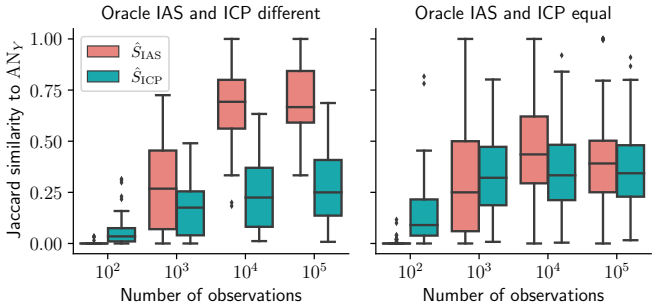


Figure E.5.: The same figure as the one presented in Fig. 4, but with weaker environments (do-interventions of strength 0.5 compared to 1 in Fig. 4). Generally, IAS performs the same for weaker interventions as for strong interventions, when there are more than 10,000 observations. Graphs represented in each boxplot: 42 (left), 58 (right).

Table E.1.: Summary of the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}_{\text{ICP}} = \emptyset)$ for weak and strong do-interventions (strength 0.5 and 1, respectively) when $d = 6$. Numbers not in parentheses are means, numbers in parentheses are medians. The level is generally unchanged when the environments have a weaker effect, but the power is lower, in the sense that the empty set is output more often.

		$\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$	$\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$	$\mathbb{P}(\hat{S}_{\text{ICP}} = \emptyset)$
Strong interventions	$n = 100$	96.6% (100%)	89.6% (98%)	52.3% (52%)
	$n = 1{,}000$	75.7% (100%)	10.0% (0%)	30.4% (14%)
	$n = 10{,}000$	83.7% (100%)	1.0% (0%)	24.9% (10%)
	$n = 100{,}000$	93.8% (100%)	0.2% (0%)	22.9% (10%)
Weak interventions	$n = 100$	99.3% (100%)	98.7% (100%)	72.0% (84%)
	$n = 1{,}000$	81.1% (100%)	40.2% (26%)	36.9% (24%)
	$n = 10{,}000$	80.8% (100%)	1.7% (0%)	27.5% (15%)
	$n = 100{,}000$	92.6% (100%)	1.1% (0%)	24.8% (14%)

E.5.6. Analysis of the Choice of q_{TB} in Section 6.3

In this section, we analyze the effect of changing the cut-off q_{TB} that determines when a gene pair is considered a true positive in Section 6.3. For the results in the main paper, we use $q_{TB} = 1\%$, meaning that the pair $(\texttt{gene}_X, \texttt{gene}_Y)$ is considered a true positive if the value of $\texttt{gene}_Y$ when intervening on $\texttt{gene}_X$ is outside of the 0.01- and 0.99-quantiles of $\texttt{gene}_Y$ in the observational distribution. In Fig. E.6, we plot the true positive rates for several other choices of q_{TB}. We compare to the true positive rate of random guessing, which also increases if the criterion becomes easier to satisfy. We observe that the choice of q_{TB} does not substantially change the excess true positive rate of our method compared to random guessing. This indicates that while the true positives in this experiments are inferred from data, the conclusions drawn in Fig. 5 are robust with respect to some modelling choices of q_{TB}.

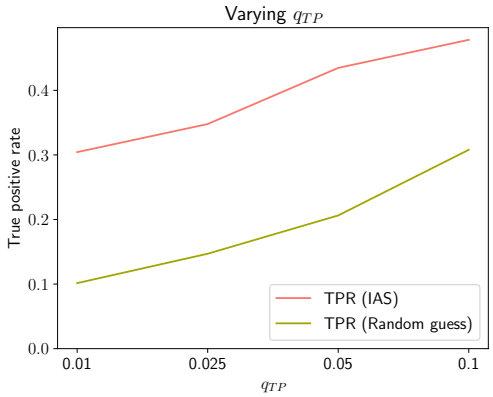


Figure E.6.: True positive rates (TPRs) for the gene experiment in Section 6.3. q_{TB} specifies the quantile in the observed distribution that an intervention effect has to exceed to be considered a true positive. While the TPR increases for our method when q_{TB} is increased, the TPR of random guessing increases comparably. This validates that changing the definition of true positives in this experiment by choosing a different q_{TB} does not change the conclusion of the experiment substantially.

E.5.7. Learning causal ancestors by estimating the I-MEC

In this section, we repeat the experiments performed in Section 6.2, this time including a procedure (here denoted $\text{IAS}_{\text{est. graph}}$), where we perform the following steps.

1. Estimate a member graph of the I-MEC and the location of the intervention sites using Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP) [Squires et al., 2020] using the implementation from the Python package CausalDAG.[1]

[1] Available at https://github.com/uhlerlab/causaldag.

2. Apply the oracle algorithm described in Section 4 to the estimated graph to obtain an estimate of $\mathcal{MI}$.

3. Output the union of all sets in the estimate of $\mathcal{MI}$.

The results for the low-dimensional experiment are displayed in Fig. E.7 and the results for the high-dimensional experiment are displayed in Table E.2. Here, we see that $\text{IAS}_{\text{est. graph}}$ generally performs well (as measured by Jaccard similarity) in the low-dimensional setting ($d = 6$), and even better than IAS for sample sizes $N \leq 10^3$, but is slightly outperformed by IAS for larger sample sizes. However, in the high-dimensional setting ($d = 100$), we observe that $\text{IAS}_{\text{est. graph}}$ fails to hold level and identifies only very few ancestors (see Table E.2). We hypothesize that the poor performance of $\text{IAS}_{\text{est. graph}}$ in the high-dimensional setting is due to $\text{IAS}_{\text{est. graph}}$ attempting to solve a more difficult task than IAS. $\text{IAS}_{\text{est. graph}}$ first estimates a full graph (here using UT-IGSP), even though only a subgraph of the full graph is of relevance in this scenario. In addition, UT-IGSP aims to estimate the site of the unknown interventions. In contrast, IAS only needs to identify nodes that are capable of blocking all paths between two variables, and does not need to know the site of the interventions.

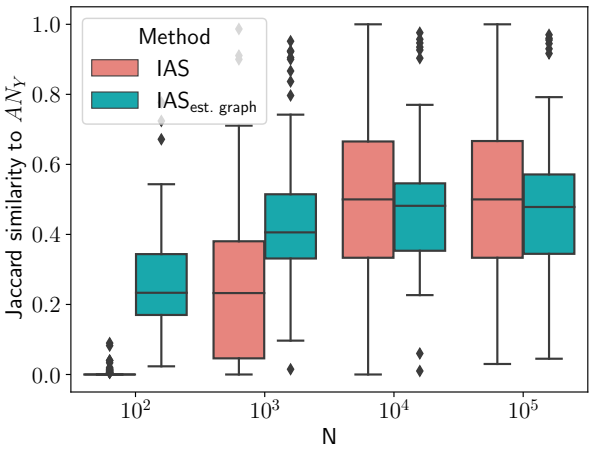


Figure E.7.: Comparison between the finite sample output of IAS and the procedure described in Appendix E.5.7, in the low-dimensional case. Generally, these procedures have similar performance, although IAS performs worse for small sample sizes but slightly better for high sample sizes.

	$d = 100, N = 10^3$		$d = 100, N = 10^4$		$d = 100, N = 10^5$	
	IAS	$\text{IAS}_{\text{est. graph}}$	IAS	$\text{IAS}_{\text{est. graph}}$	IAS	$\text{IAS}_{\text{est. graph}}$
$\mathbb{P}(S_\cdot \subseteq \text{AN}_Y)$	84.64%	15.30%	94.04%	14.92%	94.72%	14.74%
$\mathbb{P}(S_\cdot = \emptyset)$	51.96%	12.32%	12.72%	11.84%	6.98%	11.42%
$J(S_\cdot, \text{AN}_Y)$	0.19	0.10	0.33	0.10	0.35	0.11

Table E.2.: Identifying ancestors by first estimating the I-MEC of the underlying DAG and then applying the oracle algorithm of Section 4 fails to hold level and identifies fewer ancestors than applying IAS, when in a high-dimensional setting.

F. Appendix to Identifying Causal Effects using Instrumental Time Series: Nuisance IV and Correcting for the Past

F.1. Additional details for Section 2

F.1.1. Relation to VARMA Processes

In this section, we discuss that the partially observed VAR(1) process can also be viewed as a VARMA(p, q) process. In this perspective, the difficulty of identifying β when H is unobserved is linked to the non-uniqueness of vector autoregressive moving average (VARMA) process representations. The observed process $[I^\top, X^\top, Y^\top]_{t\in\mathbb{Z}}^\top$ can be obtained as a linear transformation of the VAR(1) process S and as such it has a VARMA(p, q) process representation where $p \leq d$ and $q \leq (d-1)$ [Lütkepohl, 2005, Corollary 11.1.1]. Intuitively, the dependences between I, X, Y induced by the unobserved H process can instead be modelled by serially correlated errors and higher-order memory. In contrast to a VAR process, however, the parameters of a VARMA process in standard form are not identified and different parameter settings may induce the same distribution over $[I^\top, X^\top, Y^\top]_{t\in\mathbb{Z}}^\top$ [Lütkepohl, 2005, Chapter 12.1]. As such, it is not straight-forward to obtain β from a VARMA representation of the observed process, even when choosing a canonical representation such as the echelon form or the final equations form. In this work, we propose another approach and describe how to exploit the instrumental variables idea to identify β when H is unobserved, without needing to estimate all of A.

F.1.2. Observational Equivalence

Without an instrument, the causal effect β is, due to the hidden confounding, not identifiable in general. In a fully observed Gaussian VAR(1) process, the parameter matrix (which contains the causal effect) and the covariance matrix of the noises are uniquely determined by the distribution, and can be identified by least squares regression on the previous time step, for example [Hamilton, 1994, Chapter 11]. This is not the case if parts of the system are unobserved. We consider a VAR(1) process over $H = [H^1, H^2]^\top, X$, and Y, where H is latent, and provide two different sets of parameters which entail the same observational distribution, that is, the same joint distribution of the observed process $[S_{XY,t}]_{t\in\mathbb{Z}} = [X_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$. Nevertheless, the causal effects in the two cases are different (one is 0, the other is $b \neq 0$), and so are the induced intervention distributions

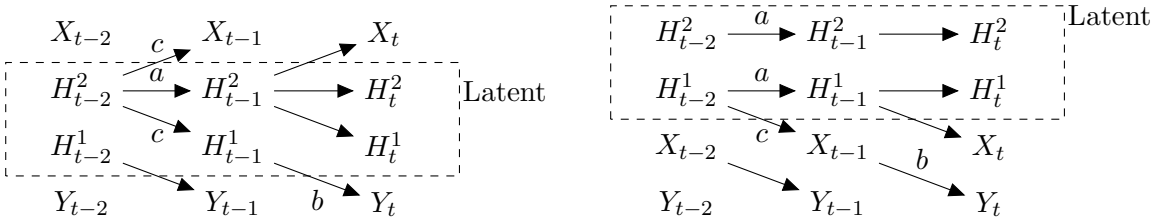


Figure F.1.: Illustration of two different causal mechanisms ($S^{(1)}$, left, and $\tilde{S^{(2)}}$, right) which are observationally equivalent for any $b, c \in \mathbb{R}$ and $a \in [-1, 1]$ and Gaussian noise distribution $\mathcal{N}(0, \mathrm{Id})$.

when intervening on X, see Appendix F.1.3 below. Consider the two coefficient matrices

$$A_1 := \begin{matrix} \\ \text{H}^1 \\ \text{H}^2 \\ \text{X} \\ \text{Y} \end{matrix} \begin{pmatrix} a & 0 & 0 & 0 \\ c & 0 & 0 & 0 \\ c & 0 & 0 & 0 \\ 0 & b & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_2 := \begin{matrix} \\ \text{H}^1 \\ \text{H}^2 \\ \text{X} \\ \text{Y} \end{matrix} \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & b & 0 \end{pmatrix}, \tag{F.1}$$

with coefficients $a \in (-1, 1)$ and $b, c \in \mathbb{R} \setminus \{0\}$. Consider the processes $S^{(1)}$ and $S^{(2)}$ satisfying

$$S^{(i)}_{t+1} = A_i S^{(i)}_t + \varepsilon^i_{t+1}, \tag{F.2}$$

with $\varepsilon^i_t \sim \mathcal{N}(0, \mathrm{Id})$. Figure F.1 depicts parts of the corresponding full time graphs. Let $S^{(1)}_{XY}$ and $S^{(2)}_{XY}$ denote the subprocesses where only X and Y are observed. The following result shows, that $S^{(1)}_{XY}$ and $S^{(2)}_{XY}$ are identically distributed, that is the two models arising from A_1 and A_2 are *observationally equivalent* [e.g., Rothenberg, 1971].

Proposition F.1. *Let $S^{(1)}, S^{(2)}$ be the processes defined from (F.2) with respective parameter matrices A_1 and A_2 from (F.1) and $\mathcal{N}(0, \mathrm{Id})$ distributed noise. Then the observed subprocesses $S^{(1)}_{XY}$ and $S^{(2)}_{XY}$ are identically distributed.*

Proof. Since each of the processes are jointly Gaussian with zero mean, their distributions are uniquely determined by the autocovariance matrices. This means that the observed processes are identically distributed if and only if $\mathbb{E}(S^{(1)}_{XY,t} {S^{(1)}_{XY,t-s}}^\top) = \mathbb{E}(S^{(2)}_{XY,t} {S^{(2)}_{XY,t-s}}^\top)$ for all $s \geq 0$. For $i \in \{1, 2\}$, the s-th autocovariance of $S^{(i)}$ is given by:

$$\mathbb{E}(S^{(i)}_t {S^{(i)}_{t-s}}^\top) = \sum_{k=0}^{\infty} A_i^{k+s} \,\mathrm{Id}\, {A_i^k}^\top.$$

Observe that

$$A_1^k = \begin{pmatrix} a^k & 0 & 0 & 0 \\ a^{k-1}c & 0 & 0 & 0 \\ a^{k-1}c & 0 & 0 & 0 \\ a^{k-2}bc & 0 & 0 & 0 \end{pmatrix} \quad A_2^k = \begin{pmatrix} a^k & 0 & 0 & 0 \\ 0 & a^k & 0 & 0 \\ 0 & a^{k-1}c & 0 & 0 \\ 0 & a^{k-2}bc & 0 & 0 \end{pmatrix}$$

for $k \geq 2$. Consequently,

$$A_1^{k+s} A_1^{k\top} = \begin{pmatrix} a^{2k+s} & a^{2k+s-1}c & a^{2k+s-1}c & a^{2k+s-2}bc \\ * & a^{2k+s-2}c^2 & a^{2k+s-2}c^2 & a^{2k+s-3}bc^2 \\ * & * & a^{2k+s-2}c^2 & a^{2k+s-3}bc^2 \\ * & * & * & a^{2k+s-4}b^2c^2 \end{pmatrix} \text{ and}$$

$$A_2^{k+s} A_2^{k\top} = \begin{pmatrix} a^{2k+s} & 0 & 0 & 0 \\ * & a^{2k+s} & a^{2k+s-1}c & a^{2k+s-2}bc \\ * & * & a^{2k+s-2}c^2 & a^{2k+s-3}bc^2 \\ * & * & * & a^{2k+s-4}b^2c^2 \end{pmatrix}$$

for $k \geq 2$ and $s \geq 0$, where the asterisks are given by symmetry of the matrices. For the case $k = 1, s = 0$, we have:

$$A_1 A_1^\top = \begin{pmatrix} a^2 & a^2c & a^2c & 0 \\ * & c^2 & c^2 & 0 \\ * & * & c^2 & 0 \\ * & * & * & b^2 \end{pmatrix} \quad \text{and} \quad A_2 A_2^\top = \begin{pmatrix} a^2 & 0 & 0 & 0 \\ * & a^2 & ac & 0 \\ * & * & c^2 & 0 \\ * & * & * & b^2 \end{pmatrix},$$

and if $k = 1, s \geq 1$:

$$A_1^{1+s} A_1^\top = \begin{pmatrix} a^{2+s} & a^{1+s}c & a^{1+s}c & 0 \\ a^{1+s}c & a^s c^2 & a^s c^2 & 0 \\ a^{1+s}c & a^s c^2 & a^s c^2 & 0 \\ a^s bc & a^{s-1}c^2 b & a^{s-1}c^2 b & 0 \end{pmatrix} \quad \text{and} \quad A_2^{1+s} A_2^\top = \begin{pmatrix} a^{2+s} & 0 & 0 & 0 \\ 0 & a^{2+s} & a^{1+s}c & 0 \\ 0 & a^{1+s}c & a^s c^2 & 0 \\ 0 & a^s bc & a^{s-1}c^2 b & 0 \end{pmatrix}.$$

For any of the above matrices M, let M_{XY} denote the 2×2 submatrix in the bottom right corner, relating to the X, Y subprocess. In all of the above cases, these coefficients relating to the X, Y subprocess coincide, that is for any $k, s \geq 0$, $(A_1^{k+s} A_1^{k\top})_{XY} = (A_2^{k+s} A_2^{k\top})_{XY}$, and since therefore $(\sum_{k=0}^\infty A_1^{k+s} A_1^{k\top})_{XY} = \sum_{k=0}^\infty (A_2^{k+s} A_2^{k\top})_{XY}$, it follows that $\mathbb{E}(S^{(1)}_{XY,t} S^{(1)\top}_{XY,t-s}) = \mathbb{E}(S^{(2)}_{XY,t} S^{(2)\top}_{XY,t-s})$ for all $s \geq 0$. □

F.1.3. Structural Causal Models and Interventions

We provide a formal introduction to structural causal models (SCMs) and interventions, which motivates the notion of a causal effect. For a more detailed introduction, see Pearl [2009].

An SCM consists of a tuple $\Pi = (\mathcal{S}, P_\varepsilon)$ where $\mathcal{S}$ is a set of structural assignments and

P_ε describes the joint distribution of the error terms. For a finite collection of variables $S^1, \dots, S^d$, with structural assignments $\mathcal{S} := \{f^1, \dots, f^d\}$ and noise distribution $P_\varepsilon = P^1 \otimes \cdots \otimes P^d$, for each $j = 1, \dots, d$, the structural equation of S^j is

$$S^j := f^j(\mathrm{PA}_j, \varepsilon^j),$$

where ε^j is distributed according to P^j and the *parents* PA_j is a subset of $\{S^1, \dots, S^d\} \setminus \{S^j\}$. The SCM induces a corresponding graph $\mathcal{G}$ over nodes $\{1, \dots, j\}$, where we draw an edge from j' to j if $S^{j'} \in \mathrm{PA}_j$. We assume that the parent sets are such that $\mathcal{G}$ is acyclic.

Similarly, we interpret the VAR process described in (1) as a structural causal model over an infinite number of nodes $[S_t]_{t\in\mathbb{Z}}$. In this case, the structural assignments are $S_t := AS_{t-1} + \varepsilon_t$, $t \in \mathbb{Z}$. Here, the error terms ε_t are assumed to be i.i.d. over time, distributed according to P_ε. Furthermore, P_ε is a product distribution, and thus the error terms are jointly independent. The SCM entails an observational distribution on the variables $[S_t]_{t\in\mathbb{Z}}$ which we denote by P_S^Π.

Formally, an intervention on an SCM is a replacement of one or more of the structural assignments at one or more time points. Such a replacement induces a new SCM that we denote by $\tilde{\Pi} = (\tilde{\mathcal{S}}, \tilde{P}_\varepsilon)$. An example of an intervention on the above VAR process is to fix the value of X for some specific time point t_0 – we write this intervention as $\mathrm{do}(X_{t_0} := x)$. Under this intervention, for $t \neq t_0$, the process still satisfies the original SCM, including assumptions on the noise variables.

The interventional distribution of Π under this intervention is defined as $P_S^{\Pi;\mathrm{do}(X_{t_0}:=x)} := P_S^{\tilde{\Pi}}$. In general, the interventional distribution of Π under an intervention is the distribution that is induced by the SCM $\tilde{\Pi}$ obtained by replacing some of the structural assignments. We require that this distribution exists and is unique. Depending on the application at hand, several interventions on the process are useful, including, for example, changing the dynamics of one component for all time points [Peters et al., 2022]. In this work, we focus on an intervention at a particular time point. When performing such an intervention $\mathrm{do}(X_{t_0} := x)$ on X_{t_0}, we have that $\frac{\partial}{\partial x}\mathbb{E}_{P_Y^{\Pi;\mathrm{do}(X_{t_0}:=x)}}[Y_{t_0+1}] = \alpha_{Y,X}$. This motivates calling $\beta = \alpha_{Y,X}$ the causal effect from X to Y. In several applications, the causal effect β is of interest by itself because it yields insight into understanding the causal structure of the problem. The causal effect, however, also comes with another benefit: it is optimal for prediction under intervention. We discuss this point of view in Section 4.3.

F.1.4. Defining Multivariate Total Causal Effects

In some cases, the effect we want to estimate may be more general than a single entry in one of the coefficient matrices A_k. In Section 2.1 we define the total causal effect of

a single variable S^i_{t-l} on S^j_t as

$$\left(\sum_{\substack{1 \leq l_1, \dots, l_m \leq p \\ l_1 + \cdots + l_m = l}} A_{l_1} \cdots A_{l_m} \right)_{j,i},$$

where A_l are the parameter matrices of the VAR(p) process. Using the method of path coefficients [Wright, 1934], we now provide a more general definition, where $\mathcal{X}$ may contain multiple variables.

Definition F.1 (Path coefficients)**.** Let S be a VAR(p) process and let $Y_t := S^{i_0}_t$ be some subprocess. Also, let $\mathcal{X}_t = [{S^{i_1}_{t-l_1}}^\top, \dots, {S^{i_m}_{t-l_m}}^\top]^\top$ be a collection of subprocesses of S. For $j = 1, \dots, m$, we define a $S^{i_j}_{t-l_j}$*-causal path* to be a directed path from $S^{i_j}_{t-l_j}$ to Y_t in the full time graph of S that does not intersect any other $S^{i_{j'}}_{t-l_{j'}}$, for $j' \neq j$. For a $S^{i_j}_{t-l_j}$-causal path $\pi : S^{i_j}_{t-l_j} \xrightarrow{e_1} \cdots \xrightarrow{e_d} S^{i_0}_t$, we define the *path coefficient* to be the product of linear coefficients along π, $c_\pi := \prod_{k=1}^d a_k$, where a_k denotes the entry in the coefficient matrix A_v, for the lag v corresponding to the edge e_k.

We can now define the total causal effect.

Definition F.2 (Total Causal Effect)**.** Let S be a VAR(p) process, let $Y_t = S^{i_0}_t$ be a subprocess of S and let $\mathcal{X} = [{S^{i_1}_{t-l_1}}^\top, \dots, {S^{i_m}_{t-l_m}}^\top]^\top$ be a collection of subprocesses of S. For $j = 1, \dots, m$, the *total causal effect*, β^j, of $S^{i_j}_{t-l_j}$ on Y_t is the sum of path coefficients c_π over all $S^{i_j}_{t-l_j}$-causal paths π from $S^{i_j}_{t-l_j}$ to Y_t, $\beta^i := \sum_{S^i_{t-l_1}\text{-causal paths } p} c_p$. Similarly the total causal effect, β, of $\mathcal{X}_t$ on Y_t is the bundling of these, $\beta := [{\beta^1}^\top, \dots, {\beta^m}^\top]^\top$.

F.2. Additional details for Section 3

F.2.1. Asymptotic variances for i.i.d. estimators

Drawing on existing results [Hall, 2005], we now provide formulas for the asymptotic variances of the NIV and CIV estimators. If a unique solution (β, α) exist to (6), the asymptotic distribution of the $\hat{\beta}_{\text{NIV},T}$ estimator in the i.i.d. setting, discussed in Section 3.2, with the weight matrix $W := \mathbb{E}[\mathcal{I}\mathcal{I}^\top]^{-1}$ (which asymptotically is optimal, see Section 3.1) is given by

$$\sqrt{T}(\hat{\beta}_{\text{NIV},T} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_1),$$

for $T \to \infty$, where the asymptotic variance Σ_1 is given by

$$\Sigma_1 := \left(\mathbb{E}(\tilde{\mathcal{X}}\mathcal{I}^\top) K^{-1} \mathbb{E}(\tilde{\mathcal{X}}\mathcal{I}^\top)^\top\right)^{-1},$$

where $\tilde{\mathcal{X}} := [\mathcal{X}^\top, \mathcal{Z}^\top]^\top$, $K = \mathbb{E}((Y - \beta\mathcal{X} - \alpha\mathcal{Z})^2)\mathbb{E}(\mathcal{I}\mathcal{I}^\top)$. Σ_1 is a $(d_\mathcal{X} + d_\mathcal{Z}) \times (d_\mathcal{X} + d_\mathcal{Z})$ matrix, with the top-left $d_\mathcal{X} \times d_\mathcal{X}$ sub-matrix describing the asymptotic variance of $\mathcal{X}$.

Similarly, the asymptotic distribution of $\hat{\beta}_{\text{CIV},T}$ in the i.i.d. setting, discussed in Section 3.1, with weight matrix $W := \mathbb{E}[\text{var}(\mathcal{I}|\mathcal{B})]^{-1}$ is

$$\sqrt{T}(\hat{\beta}_{\text{CIV},T} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_2),$$

where

$$\Sigma_2 := \left(\mathbb{E}[\text{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})]K^{-1}\mathbb{E}[\text{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})^\top]\right)^{-1},$$

and $K = \mathbb{E}[\text{var}((Y - \beta\mathcal{X})^2|\mathcal{B})]\mathbb{E}[\text{var}(\mathcal{I}|\mathcal{B})]$.

F.2.2. Asymptotic variances for estimators in time series

Closed-form expressions for the asymptotic variances of the NIV and CIV estimators can also be found for the VAR process presented in Section 4, but these are slightly more involved than for the i.i.d. setting presented in Appendix F.2.1. Assume the setting as described in Theorem 5 and consider the NIV estimator $\hat{\beta}_{\text{NIV},T}$. We then have

$$\sqrt{T}(\hat{\beta}_{\text{NIV},T} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_1),$$

where the asymptotic variance Σ_1 is given by

$$\Sigma_1 := \left(\mathbb{E}(\tilde{\mathcal{X}}\mathcal{I}^\top)K^{-1}\mathbb{E}(\tilde{\mathcal{X}}\mathcal{I}^\top)^\top\right)^{-1},$$

where $\tilde{\mathcal{X}} := [X_{t-1}^\top, \quad Y_{t-1}^\top]^\top, \mathcal{I} := \{I_{t-2}, ..., I_{t-m-1}\}$ and $K := \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{(T)}}\sum_{t=1}^{T}(Y_t - \beta\mathcal{X} - \alpha\mathcal{Z})\mathcal{I}^\top\right)$ using the optimal choice of weight matrix, $W = K^{-1}$, see Hall [2005, Chapter 3].

Now for the CIV estimator, see Theorem 4, the asymptotic distribution of $\hat{\beta}_{\text{CIV},T}$ is

$$\sqrt{T}(\hat{\beta}_{\text{CIV},T} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_2),$$

where

$$\Sigma_2 := \left(\mathbb{E}[\text{cov}(X_{t-1}, I_{t-2}|\mathcal{B}_t)]K^{-1}\mathbb{E}[\text{cov}(X_{t-1}, I_{t-2}|\mathcal{B}_t)^\top]\right)^{-1},$$

and $K := \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{(T)}}\sum_{t=1}^{T}(r_{Y_t} - \beta r_{X_{t-1}})r_I^\top\right)$ with $r_i := i - \mathbb{E}[i|\mathcal{B}_t]$ using the optimal choice of weight matrix, $W = K^{-1}$, see Hall [2005, Chapter 3].

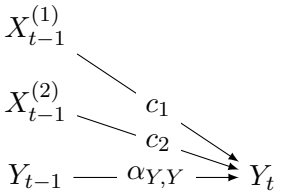


Figure F.2.: Subgraph of the full-time graph of the structure discussed in Example F.1 when $d_X = 2$. For simplicity, we don't draw nodes corresponding to the instrument, I, and the confounder, H.

F.3. Additional details for Section 4

F.3.1. Example of a distribution that does not satisfy the rank requirement in Theorem 6

In Section 4.2.2, we have developed a criterion for identifiability that depends on the parameter matrix of a process satisfying Assumption (A2). We have showed in Corollary 1 that if parameter matrices are drawn from a distribution with density with respect to Lebesgue measure, then the identifiability criterion holds almost surely. In this section, we provide an example of a parameter matrix that does not satisfy the criterion.

Example F.1. *Consider the case where $d_X > 1, d_I = 1$, and $\alpha_{X,X} = \mathrm{diag}(c, \dots, c)$ for a $c \in \mathbb{R}$. By Theorem 6, β is not identifiable by NIV: this follows because A_{XY} is a lower triangular matrix with $c, \dots, c$ (d_X times) and $\alpha_{Y,Y}$ on the diagonal, and the Jordan form J is a diagonal matrix with the same diagonal entries. Hence there are d_X Jordan blocks with the same eigenvalue c so the causal effect β is not identified by NIV. On the contrary, when $\alpha_{X,X} = \mathrm{diag}(c_1, \dots, c_{d_X})$ (see Fig. F.2) where $c_i \neq c_j$ for all $i \neq j$, β is identified by NIV if also $\alpha_{Y,Y} \neq c_i$ for all i.*

Example F.2. *In the case of $d_X = 1, d_I = 1$, and $\alpha_{X,I} \neq 0$, β is identifiable. If, for example, $\alpha_{Y,Y} = \alpha_{X,X} =: \alpha$, we have,*

$$A_{XY} = \begin{pmatrix} \alpha & 0 \\ \beta & \alpha \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ \beta & 0 \end{pmatrix}}_{=:P} \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix} \underbrace{\begin{pmatrix} 0 & \frac{1}{\beta} \\ 1 & 0 \end{pmatrix}}_{=P^{-1}}.$$

This has only one Jordan block with algebraic multiplicity $m = 2$ and

$$\left(P^{-1} \begin{pmatrix} \alpha_{X,I} \\ 0 \end{pmatrix} \right)_m = [1, 0] \begin{pmatrix} \alpha_{X,I} \\ 0 \end{pmatrix} = \alpha_{X,I} \neq 0,$$

where $(\cdot)_m$ refers to the m-th entrance, so by Theorem 6, β is identifiable (and by similar arguments the same holds if $\alpha_{X,X} \neq \alpha_{Y,Y}$.

Algorithm F.1 Linear prediction under the intervention $\mathrm{do}(X_t := x)$

Input: Causal parameter β, sample $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}] \in \mathbb{R}^{d_X \times (t-1)}$, $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_t] \in \mathbb{R}^{1 \times t}$, interventional value x, lag parameters m and l

1: Compute the residual process $r_s := \mathbf{Y}_{s+1} - \beta \mathbf{X}_s$ for $s = 1, \ldots, t-1$.
2: For $s > \max(k, m)$, linearly regress r_s on $\{\mathbf{X}_{s-k}, k = 1, \ldots, m\}$ and $\{\mathbf{Y}_{s-j}, j = 0, \ldots, l\}$ to obtain coefficients $\hat{\alpha}^k_{Y,X}$ and $\hat{\alpha}^j_{Y,Y}$.

Output:
3: Prediction $\hat{\mathbf{Y}}_{t+1} = \beta x + \sum_{k=1}^{m} \hat{\alpha}^k_{Y,X} \mathbf{X}_{t-k} + \sum_{j=0}^{l} \hat{\alpha}^j_{Y,Y} \mathbf{Y}_{t-j}$.

F.3.2. Algorithm for Prediction under Interventions

This section contains Algorithm F.1, the procedure for using a causal effect to predict under interventions, as discussed in Section 4.3.

F.4. Proofs

F.4.1. Proof of Conditional IV

For completeness, we now prove our statement about CIV from Section 3.1.

Proposition F.2. *Consider a linear SCM (see Appendix F.1.3) over variables V, and let $\mathcal{I}, \mathcal{X}, \mathcal{B}, \{Y\} \subseteq V$ be disjoint collections of variables from V, and let $\mathcal{G}$ be the corresponding DAG. Assume that $\mathcal{I}, \mathcal{X}$ and Y have zero mean and finite second moment and let β be the causal coefficient with which $\mathcal{X}$ enters the structural equation for Y (some of the entries of β can be zero, so not all variables in $\mathcal{X}$ have to be parents of Y). We consider the following three requirements on $\mathcal{I}, \mathcal{X}, \mathcal{B}$ and Y*

(CIV1) *$\mathcal{I}$ and Y are d-separated given $\mathcal{B}$ in the graph $\mathcal{G}_{\mathcal{X} \not\to Y}$, that is the graph $\mathcal{G}$ where all direct edges from $\mathcal{X}$ to Y are removed,*

(CIV2) *$\mathcal{B}$ is not a descendant of $\mathcal{X} \cup Y$ in $\mathcal{G}$, and*

(CIV3) *the matrix $\mathbb{E}[\mathrm{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})]$ has rank $d_{\mathcal{X}}$, that is, full row rank.*

If requirements (CIV1) and (CIV2) are met, $Y - \beta\mathcal{X} \perp\!\!\!\perp \mathcal{I}|\mathcal{B}$, and in particular β satisfies the CIV moment equation

$$\mathbb{E}[\mathrm{cov}(Y - \beta\mathcal{X}, \mathcal{I}|\mathcal{B})] = 0. \tag{F.3}$$

If, additionally, requirement (CIV3) is met, β is the unique solution to this equation,

$$\mathbb{E}[\mathrm{cov}(Y - b\mathcal{X}, \mathcal{I}|\mathcal{B})] = 0 \implies b = \beta.$$

Proof. Due to the additive, linear structure of the SCM, we can write Y as a

$$Y = \beta\mathcal{X} + \pi\mathcal{B} + \gamma R + \varepsilon^Y,$$

where R are those parents of Y that are not in $\mathcal{X} \cup \mathcal{B}$ and $\pi \in \mathbb{R}^{d_Y \times d_\mathcal{B}}, \gamma \in \mathbb{R}^{d_Y \times d_R}$ are some coefficients.

We claim (1) that any path from $\mathcal{I}$ to R is blocked by $\mathcal{B}$ and (2) that $\mathbb{E}[\mathrm{cov}(\varepsilon^Y, \mathcal{I}|\mathcal{B})] = 0$. It then follows from the global Markov property [Lauritzen, 1996], that $\mathbb{E}[\mathrm{cov}(R, \mathcal{I}|\mathcal{B})] = 0$, and trivially also $\mathbb{E}[\mathrm{cov}(\mathcal{B}, \mathcal{I}|\mathcal{B})] = 0$. Hence, since $Y - \beta\mathcal{X} = \pi\mathcal{B} + \gamma R + \varepsilon^Y$, it follows that $\mathbb{E}[\mathrm{cov}(Y - \beta\mathcal{X}, \mathcal{I}|\mathcal{B})] = \mathbb{E}[\mathrm{cov}(\pi\mathcal{B} + \gamma R + \varepsilon^Y, \mathcal{I}|\mathcal{B})] = 0$.

For (1), suppose for a contradiction that a path π between $\mathcal{I}$ and R that is unblocked given $\mathcal{B}$ exists. Case 1: π does not contain any edge from $\mathcal{X}$ to Y. Then, the path that concatenates π with the corresponding edge from R to Y is an unblocked path (given $\mathcal{B}$) in the graph, where the edges from $\mathcal{X}$ to Y are removed. This contradicts requirement (CIV1). Case 2: π contains an edge from $X \in \mathcal{X}$ to Y. Then, π contains either the structure $X \to Y \leftarrow$ or the structure $X \to Y \to$. The first case implies $\mathcal{B} \cap \mathrm{DE}\, Y \neq \emptyset$, violating requirement (CIV2). In the second case, we either have that there is a directed path from Y to $\mathcal{I}$ that is unblocked by $\mathcal{B}$, violating requirement (CIV1) or, again that $\mathcal{B} \cap \mathrm{DE}\, Y \neq \emptyset$, violating requirement (CIV2).

For (2), we have that neither $\mathcal{B}$ nor $\mathcal{I}$ are descendants of Y: $\mathcal{B}$ cannot be a descendant of Y due to requirement (CIV2), and by requirement (CIV1), $\mathcal{I}$ can only be a descendant if $\mathcal{B}$ is also a descendant, which is not possible. Every variable in the linear SCM can be rewritten as a function only of noise terms corresponding to ancestors. Applying this to $\mathcal{B}$ and I, we have (since Y is not an ancestor of neither $\mathcal{I}$ nor $\mathcal{B}$ and because ε^Y is independent of all other noise terms in the SCM) that ε^Y is independent of $(\mathcal{B}, \mathcal{I})$ and it follows that $\mathbb{E}[\mathrm{cov}(\varepsilon^Y, \mathcal{I}|\mathcal{B})] = 0$.

When we additionally assume requirement (CIV3), the solution to (F.3) is unique because the equation can be rewritten to

$$b\mathbb{E}[\mathrm{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})] = \mathbb{E}[\mathrm{cov}(Y, \mathcal{I}|\mathcal{B})],$$

and by the assumption of full row rank of $\mathbb{E}[\mathrm{cov}(\mathcal{X}, \mathcal{I}|\mathcal{B})]$, this can have at most one solution. □

F.4.2. Proof of Theorem 1

Theorem 1. *Consider a time series S generated according to Assumption (A1), and finite disjoint collections A, B, C. If A and C are d-separated given B in $\mathcal{G}_{full}$ then $A \perp\!\!\!\perp C|B$.*

Proof. Our proof is inspired by Lauritzen et al. [1990, Sec. 6]. Let us write $\mathcal{G}_{[s,t]}$ for the subgraph of a time series (sub)graph $\mathcal{G}$, where only vertices $V_{[s,t]} := \{S_v^i | i = 1, \ldots, d, s \leq v \leq t\}$ are included. Let further s_0, t_0 be the largest and smallest time points respectively such that $A \cup C \cup B \subseteq V_{[s_0,t_0]}$. Let $q \in \mathbb{N}_{>p}$ such that if two nodes in $V_{[s_0,t_0]}$ are d-connected (with empty conditioning set) in $\mathcal{G}_{\text{full}}$, then there is a d-connecting path in

$\mathcal{G}_{[s_0-q,t_0]}$. Define the set

$$\begin{aligned}\mathcal{A} := \{n \in \mathbb{N} \,|\, &\text{for all } \mathcal{G}^* \text{ that are graphs over nodes with time indices between } s_0-q \text{ and } t_0 \text{ s.t.}\\ &\text{all edges in } \mathcal{G}^* \text{ point 'forward in time', i.e., } \forall k \in \mathbb{N}, \text{ there is no edge } S_t^i \to S_{t-k}^j \text{ and}\\ &|V^+| = n, \text{ where } V^+ := V_{[s_0,t_0]} \text{ and}\\ &\text{for all VAR processes whose structure is specified by } \mathcal{G}^0 := \mathcal{G}^*_{[s_0-q,s_0-1]} \text{ and}\\ &\text{for all } A^*, B, C^* \subseteq \mathcal{G}^* \text{ such that}\\ &AN_{\mathcal{G}^+}(A^+ \cup B \cup C^+) = V^+ (\text{where } A^+ := A^*_{[s_0,t_0]}, C^+ := C^*_{[s_0,t_0]}, \mathcal{G}^+ := \mathcal{G}^*_{[s_0,t_0]}), \text{ and}\\ &A^* = A^+ \cup (\mathrm{PA}_{\mathcal{G}^*}(A^+) \cap V^0) \text{ and } C^* = C^+ \cup (\mathrm{PA}_{\mathcal{G}^*}(C^+) \cap V^0)\\ &\text{where } V^0 \text{ are the nodes of } \mathcal{G}^0\\ &\text{we have}\\ &A^* \perp_{\mathcal{G}^*} C^* | B \;\Rightarrow\; A^* \perp\!\!\!\perp C^* | B\},\end{aligned}$$

where $\perp_{\mathcal{G}}$ indicates d-separation in $\mathcal{G}$.

We show below, by induction, that $\mathcal{A} = \mathbb{N}$. This suffices to prove the statement of the theorem because of the following line of arguments. Let $V^0 := V_{[s_0-q,s_0-1]}$ and $V^+ := AN_{\mathcal{G}_{\text{full}}}(A \cup C \cup B)_{[s_0,t_0]}$. Let $\mathcal{G}^*$ be the graph $\mathcal{G}_{\text{full}}$ restricted to the nodes in $V^* := V^0 \cup V^+$. Then, $A \perp_{\mathcal{G}^*} C|B$ (as $\mathcal{G}^*$ is a subgraph of $\mathcal{G}_{\text{full}}$). If $V^+ \neq (A \cup C \cup B)$, we enlarge A and C to the disjoint sets A^+ and C^+ such that $A^+ \perp_{\mathcal{G}^*} C^+|B$ and $V^+ = A^+ \cup B \cup C^+$. Let us define $A^* := A^+ \cup (\mathrm{PA}_{\mathcal{G}^*}(A^+) \cap V^0)$ and $C^* := C^+ \cup (\mathrm{PA}_{\mathcal{G}^*}(C^+) \cap V^0)$. Importantly, these two sets are disjoint (otherwise, A^+ and C^+ would have a joint parent not in B, violating $A^+ \perp_{\mathcal{G}^*} C^+|B$). We then have that $A^* \perp_{\mathcal{G}^*} C^*|B$ (Indeed, if there is an open path from a node in $a \in A^*$ to a node in $c \in C^*$, given B, then there is an open path between a node in A^+ (either a itself or its child in A^+) to a node in C^+ (either c itself or its child in C^+), violating $A^+ \perp_{\mathcal{G}^*} C^+|B$). But then $\mathcal{A} = \mathbb{N}$ implies $A^* \perp\!\!\!\perp C^*|B$. And this implies that $A \perp\!\!\!\perp C|B$, as A and B are subsets of A^* and B^*, respectively.

Let us now prove that $\mathcal{A} = \mathbb{N}$ by, (1), proving $1 \in \mathcal{A}$ and $2 \in \mathcal{A}$ and, (2), proving $n \in \mathcal{A}$ implies $n+1 \in \mathcal{A}$.

(1) We now prove that $1 \in \mathcal{A}$ and $2 \in \mathcal{A}$.
The only non-trivial statement occurs when $A^* = \{a\} \neq \emptyset$ and $C^* = \{c\} \neq \emptyset$ and $B = \emptyset$. Because $A^* \perp_{\mathcal{G}^*} C^*$, we have $AN_{\mathcal{G}^*}(A^*) \cap AN_{\mathcal{G}^*}(C^*) = \emptyset$. This implies $AN_{\mathcal{G}_{\text{full}}}(A^*) \cap AN_{\mathcal{G}_{\text{full}}}(C^*) = \emptyset$ because of the repetitive structure in a full time graph and the choice of q.

(2) We now prove that $n \in \mathcal{A}$ implies $n+1 \in \mathcal{A}$.
Assume $n \in \mathcal{A}$ and consider $\mathcal{G}^*$, $\mathcal{G}^0$, $\mathcal{G}^+$, V^+, A^*, A^+, B, C^*, C^+ as described in set $\mathcal{A}$ with $|V^+| = n+1$. Consider a node $\lambda \in V^+$ that is a sink node in $\mathcal{G}^*$.

First, assume that $\lambda \in A^+$. Then $\mathrm{PA}_{\mathcal{G}^*}(\lambda) \subseteq (A^* \setminus \{\lambda\}) \cup B$ (because d-separation would be violated if $C^* \cap \mathrm{PA}(\lambda) \neq \emptyset$). Thus, it follows that

$$\lambda \perp\!\!\!\perp C^* | B \cup (A^* \setminus \{\lambda\}). \tag{F.4}$$

(Indeed, $\mathrm{PA}_{\mathcal{G}_{\text{full}}}(\lambda) = \mathrm{PA}_{\mathcal{G}^*}(\lambda)$ and thus there exist a coefficient vector $\gamma \in \mathbb{R}^{|\mathrm{PA}_{\mathcal{G}^*}(\lambda)|}$ such that $\lambda = \gamma^\top \mathrm{PA}_{\mathcal{G}^*}(\lambda) + \varepsilon^\lambda$; it then follows from the $MA(\infty)$ representation of S [Hamilton, 1994], that $\lambda \perp\!\!\!\perp C^* \cup (B \cup (A^* \setminus \{\lambda\}) \setminus \mathrm{PA}_{\mathcal{G}^*}(\lambda)) \mid \mathrm{PA}_{\mathcal{G}^*}(\lambda)$; the claimed independence then follows with the weak union property.) Further, $A^* \setminus \{\lambda\} \perp_{\mathcal{G}^{*m}} C^* | B$, where $\mathcal{G}^m$ denotes moralization of graph $\mathcal{G}$ [Lauritzen, 1996], as d-separation is equivalent to separation in the moralized graph. But then, $A^* \setminus \{\lambda\} \perp_{(\mathcal{G}^{*m})_{V^* \setminus \{\lambda\}}} C^* | B$, as this graph contains no more edges. And therefore, $(A^* \setminus \{\lambda\}) \perp_{(\mathcal{G}^*_{V^* \setminus \{\lambda\}})^m} C^* | B$ as, again, the graph contains no more edges. By the induction hypothesis $n \in \mathcal{A}$ and thus

$$A^* \setminus \{\lambda\} \perp\!\!\!\perp C^* | B. \tag{F.5}$$

Combining (F.4) and (F.5) by the contraction property, it follows that

$$A^* \perp\!\!\!\perp C^* | B.$$

Second, assume that $\lambda \in C^+$. The argument follows in the same way as in the case $\lambda \in A^+$.

Third, assume that $\lambda \in B$ (these are all cases since $A^+ \cup B \cup C^+ = V^+$). Since B separates A^* and C^* in $(\mathcal{G}^*)^m$, then $B \setminus \{\lambda\}$ also separates A^* and C^* in $(\mathcal{G}^{*m})_{V^* \setminus \{\lambda\}}$ (as it has no more edges), and therefore $B \setminus \{\lambda\}$ also separates A^* and C^* in $(\mathcal{G}^*_{V^* \setminus \{\lambda\}})^m$, since this graph, again, has no more edges. By the induction hypothesis $n \in \mathcal{A}$, so

$$A^* \perp\!\!\!\perp C^* | (B \setminus \{\lambda\}). \tag{F.6}$$

We now prove a second independence statement. We now make a case distinction (a) Assume that

$$\mathrm{PA}_{\mathcal{G}^+}(\lambda) \cap A^* \neq \emptyset \quad \text{or} \quad AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}) \cap AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(A^*)_{[s_0-q,s_0-1]}) \neq \emptyset.$$

Then, it follows that

$$\mathrm{PA}_{\mathcal{G}^+}(\lambda) \cap C^* = \emptyset \quad \text{and } AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}) \cap AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(C^*)_{[s_0-q,s_0-1]}) = \emptyset. \tag{F.7}$$

(Indeed, if the statement on the left-hand side would be false, then there is a d-connecting path between A^* and C^*, given B: this goes from the element in $\mathrm{PA}_{\mathcal{G}^+} \cap C^*$ to λ (which is in B) and then either to the element in $\mathrm{PA}_{\mathcal{G}^+} \cap A^*$ or to the common ancestor of $\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}$ and $\mathrm{PA}_{\mathcal{G}^*}(A^*)_{[s_0-q,s_0-1]}$ and then to the corresponding element in A^*. If the statement on the right-hand side would be false, then we can use the same path but this time going via the common ancestor of $\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}$ and $\mathrm{PA}_{\mathcal{G}^*}(C^*)_{[s_0-q,s_0-1]}$.)

But then it follows that

$$\lambda \perp\!\!\!\perp C^* | A^* \cup (B \setminus \{\lambda\}). \tag{F.8}$$

(Indeed, noting that $\mathrm{PA}_{\mathcal{G}^+}(\lambda) \subseteq A^* \cup B \setminus \{\lambda\}$, we can replace the left-hand side by the MA(∞) representation of $\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}$. For C^*, we repeatedly

use the structural equations except for variables in $B \setminus \{\lambda\}$ or variables in $\mathrm{PA}_{\mathcal{G}^*}(C^*)_{[s_0-q,s_0-1]}$ (other variables will not occur: If there was a variable in A^*, for example, there would be a directed path from A^* to C^*). We then use the $\mathrm{MA}(\infty)$ representation of $\mathrm{PA}_{\mathcal{G}^*}(C^*)_{[s_0-q,s_0-1]}$. The statement then follows from the fact that $\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}$ and $\mathrm{PA}_{\mathcal{G}^*}(C^*)_{[s_0-q,s_0-1]}$ do not have common ancestors, see (F.7).

Combining (F.6) and (F.8) using the contraction property, it follows that $C^* \perp\!\!\!\perp (\{\lambda\} \cup A^*)|(B \setminus \{\lambda\})$, and by the weak union property that

$$C^* \perp\!\!\!\perp A^*|B.$$

(b) Now assume that

$$\mathrm{PA}_{\mathcal{G}^+}(\lambda) \cap A^* = \emptyset \quad \text{and} \quad AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(\lambda)_{[s_0-q,s_0-1]}) \cap AN_{\mathcal{G}^*}(\mathrm{PA}_{\mathcal{G}^*}(A^*)_{[s_0-q,s_0-1]}) = \emptyset.$$

Similarly as in case (a) it follows that

$$\lambda \perp\!\!\!\perp A^*|C^* \cup (B \setminus \{\lambda\}). \tag{F.9}$$

Combining (F.6) and (F.9) using the contraction property, it follows that $(\{\lambda\} \cup C^*) \perp\!\!\!\perp A^*|(B \setminus \{\lambda\})$, and by the weak union property that

$$C^* \perp\!\!\!\perp A^*|B.$$

This concludes the proof. □

F.4.3. Proof of Theorem 2

Theorem 2 (Nuisance IV)**.** *Consider a linear SCM (see Appendix F.1.3) over variables V, and let $\mathcal{I}$, $\mathcal{X}$, $\mathcal{Z}$, $\mathcal{B}$, $\{Y\} \subseteq V$ be disjoint collections of variables from V, and let $\mathcal{G}$ be the corresponding DAG. Assume that $\mathcal{I}, \mathcal{X}, \mathcal{Z}$ and Y have zero mean and finite second moment and let β and α be the causal coefficients with which $\mathcal{X}$ and $\mathcal{Z}$ enter the structural equation for Y, respectively (some of the entries of β and α can be zero, so not all variables in $\mathcal{X}$ and $\mathcal{Z}$ have to be parents of Y). Let $\tilde{\mathcal{X}} := \mathcal{X} \cup \mathcal{Z}$. If requirements (CIV1) to (CIV3) are satisfied in $\mathcal{G}$ for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$ and Y, the causal effect β of $\mathcal{X}$ on Y is identified by $\mathrm{NIV}_{\mathcal{X}\to Y}(\mathcal{I}, \mathcal{Z}|\mathcal{B})$.*

Proof. By satisfaction of requirements (CIV1) to (CIV3), the causal effect $\tilde{\beta} = [\beta, \alpha]$ of $\tilde{\mathcal{X}} = \mathcal{X} \cup \mathcal{Z}$ on Y is identified by the instrument $\mathcal{I}$ and the conditioning set $\mathcal{B}$ by Proposition F.2 in Appendix F.4.1. In particular, also the sub-vector of the IV estimate corresponding to $\mathcal{X}$ is identified. □

F.4.4. Proof of Proposition 1

Proposition 1. *If an effect can be identified by CIV and by NIV, then the estimators cannot be strictly sorted in terms of asymptotic variance. More specifically, there exist*

data generating processes, for which CIV has strictly smaller asymptotic variance and others, for which NIV has strictly smaller asymptotic variance.

Proof. We show this by considering two SCMs over 6 variables $S = [H, I, X, Y, Z, B]$ given by $S \coloneqq AS + \varepsilon$ where A is such that the resulting graph is acyclic and admits the graphical model in Fig. 3 (right) and $\varepsilon \sim \mathcal{N}(0, \Gamma)$; we provide two concrete choices for A and Γ below. We consider both the $\mathrm{CIV}_{X\to Y}(I|B)$ and the $\mathrm{NIV}_{X\to Y}([I, B], Z)$ estimates of β, the causal effect of X on Y, and provide two sets of parameters (A^I, Γ^I) and (A^{II}, Γ^{II}) such that if X is generated according to (A^I, Γ^I), the CIV estimator has a lower asymptotic variance than the NIV estimator, and if S is generated according to (A^{II}, Γ^{II}), the CIV estimator has a higher asymptotic variance than the NIV estimator.

$$A^I \coloneqq \begin{array}{c} \\ \text{H} \\ \text{I} \\ \text{X} \\ \text{Y} \\ \text{Z} \\ \text{B} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{H} & \text{I} & \text{X} & \text{Y} & \text{Z} & \text{B} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.185 \\ 21.095 & 6.885 & 0 & 0 & 0 & -5.969 \\ -7.244 & 0 & 16.499 & 0 & -1.892 & 0 \\ 1.921 & 0 & 0 & 0 & 0 & 2.62 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\Gamma^I \coloneqq \begin{array}{c} \\ \text{H} \\ \text{I} \\ \text{X} \\ \text{Y} \\ \text{Z} \\ \text{B} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{H} & \text{I} & \text{X} & \text{Y} & \text{Z} & \text{B} \end{array} \\ \begin{pmatrix} 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 \end{pmatrix} \end{array},$$

and

$$A^{II} \coloneqq \begin{array}{c} \\ \text{H} \\ \text{I} \\ \text{X} \\ \text{Y} \\ \text{Z} \\ \text{B} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{H} & \text{I} & \text{X} & \text{Y} & \text{Z} & \text{B} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2.918 \\ -22.439 & 3.519 & 0 & 0 & 0 & 4.282 \\ 19.964 & 0 & 4.737 & 0 & 4.011 & 0 \\ 0.884 & 0 & 0 & 0 & 0 & -7.97 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\Gamma^{II} \coloneqq \begin{array}{c} \\ \text{H} \\ \text{I} \\ \text{X} \\ \text{Y} \\ \text{Z} \\ \text{B} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{H} & \text{I} & \text{X} & \text{Y} & \text{Z} & \text{B} \end{array} \\ \begin{pmatrix} 3.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.2 \end{pmatrix} \end{array}.$$

We can now use the formulas from Appendix F.2.1 to get the asymptotic variances, using that $\mathbb{E}[S] = 0$ and $\mathbb{E}[SS^\top] = (1 - A)^{-1}\Gamma(1 - A)^{-\top}$. When the data generating mechanism is (A^I, Γ^I), the asymptotic distributions are specified by $\sqrt{T}(\hat{\beta}_{\mathrm{CIV}} - \beta) \sim \mathcal{N}(0, 524.4)$ and $\sqrt{T}(\hat{\beta}_{\mathrm{NIV}} - \beta) \sim \mathcal{N}(0, 522.7)$. Furthermore, when the data generating mechanism is (A^{II}, Γ^{II}), the asymptotic distributions are $\sqrt{T}(\hat{\beta}_{\mathrm{CIV}} - \beta) \sim \mathcal{N}(0, 320.0)$

and $\sqrt{T}(\hat{\beta}_{\text{NIV}} - \beta) \sim \mathcal{N}(0, 575.4)$, respectively. □

F.4.5. Proof of Theorem 3

Theorem 3 (Time series IV by marginalization)**.** *Consider a process $S = [S_t]_{t\in\mathbb{Z}}$ satisfying Assumption (A1) with full time graph $\mathcal{G}_{full}$. Let Y be some node in $\mathcal{G}_{full}$ and let $\mathcal{X}, \mathcal{I}, \mathcal{Z}$, and $\mathcal{B}$ be disjoint collections of nodes from $\mathcal{G}_{full}$. Let $\tilde{\mathcal{X}} := \mathcal{X} \cup \mathcal{Z}$ and define $M := \{Y\} \cup \mathcal{X} \cup \mathcal{I} \cup \mathcal{Z} \cup \mathcal{B}$. Assume that requirements (CIV1') and (CIV2) are satisfied for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$ and Y in $\mathcal{G}_M$ (see Definition 1). Then, the following three statements hold. (i) The total causal effect $[\beta, \alpha]$ of $[\mathcal{X}^\top, \mathcal{Z}^\top]^\top$ on Y satisfies the NIV moment equation*

$$\mathbb{E}[\text{cov}(Y - b\mathcal{X} - a\mathcal{Z}, \mathcal{I}|\mathcal{B})] = 0. \tag{8}$$

(ii) Further, if requirement (CIV3) is satisfied for $\mathcal{I}, \tilde{\mathcal{X}}, \mathcal{B}$, then $[\beta, \alpha]$ is the unique solution to (8). (iii) If, additionally, $\mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{Z}$ and $\mathbf{B}$ are observations of $\mathcal{X}, Y, \mathcal{I}, \mathcal{Z}$ and $\mathcal{B}$ at T time points, W is a positive definite matrix, and

$$[\hat{b}, \hat{a}] := \arg\min_{b,a} \|\hat{\text{cov}}(\mathbf{Y} - b\mathbf{X} - a\mathbf{Z}, \mathbf{I}|\mathbf{B})\|_W^2, \tag{9}$$

then $\hat{b}$ is a consistent estimator for β.

Proof. We first show part (i). Due to the additive, linear structure in Assumption (A1), we can rewrite Y as a linear combination of the parents of Y in $\mathcal{G}_{\text{full}}$ plus some additive noise, $Y = a_1 p_1 + \ldots + a_m p_m + \varepsilon^Y$, where $a_1, \ldots, a_m$ are coefficients and $p_1, \ldots, p_m$ are nodes from $\mathcal{G}_{\text{full}}$. Similarly, we can recursively decompose the parents into their parents (in $\mathcal{G}_{\text{full}}$) and noise without replacing variables in $\mathcal{X} \cup \mathcal{Z} \cup \text{ND}(\mathcal{X} \cup \mathcal{Z}) \cup \mathcal{B}$, until the first time we have a decomposition

$$Y = \beta\mathcal{X} + \alpha\mathcal{Z} + \pi\mathcal{B} + \gamma R + \varepsilon,$$

where R are d_R variables from $\mathcal{G}_{\text{full}}$ that are not in $\mathcal{B}$ or on any path from $\mathcal{X} \cup \mathcal{Z}$ to Y in $\mathcal{G}_{\text{full}}$,

$$R \cap (\mathcal{B} \cup \mathcal{X} \cup \mathcal{Z} \cup \text{DE}\,\mathcal{X} \cup \mathcal{Z}) = \emptyset,$$

(R may include descendants of $\mathcal{B}$) and ε are all the weighted noise variables accumulated when doing the decomposition in parents, β and α are the total causal effects of $\mathcal{X}$ and $\mathcal{Z}$ on Y, respectively, for some coefficients $\pi \in \mathbb{R}^{1\times d_\mathcal{B}}, \gamma \in \mathbb{R}^{1\times d_R}$. The coefficients in front of $\mathcal{X}$ and $\mathcal{Z}$ are indeed β and α, because the total causal effect is the product along all paths from $\mathcal{X} \cup \mathcal{Z}$ to Y (see Appendix F.1.4), and by the assumption that $\mathcal{B} \subseteq \text{ND}(\mathcal{X} \cup \mathcal{Z})$, no direct path from $\mathcal{X} \cup \mathcal{Z}$ to Y is blocked by $\mathcal{B}$.

We claim (1) that any path in $\mathcal{G}_{\text{full}}$ from $\mathcal{I}$ to R is blocked by $\mathcal{B}$ and (2) that $\mathbb{E}[\text{cov}(\varepsilon, \mathcal{I}|\mathcal{B})] = 0$. It then follows from Theorem 1, that $\mathbb{E}[\text{cov}(R, \mathcal{I}|\mathcal{B})] = 0$, and trivially also $\mathbb{E}[\text{cov}(\mathcal{B}, \mathcal{I}|\mathcal{B})] = 0$. Hence, since $Y - \beta\mathcal{X} - \alpha\mathcal{Z} = \pi\mathcal{B} + \gamma R + \varepsilon$, it follows that $\mathbb{E}[\text{cov}(Y - \beta\mathcal{X} - \alpha\mathcal{Z}, \mathcal{I}|\mathcal{B})] = \mathbb{E}[\text{cov}(\pi\mathcal{B} + \gamma R + \varepsilon, \mathcal{I}|\mathcal{B})] = 0$.

For (1), suppose for a contradiction that there exist $i \in \mathcal{I}$ and $m \in R$ and a path

$p : i - v_1 - \cdots - v_n - m$ in $\mathcal{G}_{\text{full}}$ that is unblocked given $\mathcal{B}$, where $v_1 - v_2$ indicates a directed edge that can have any orientation. Since p is unblocked given $\mathcal{B}$, this indicates that the non-colliders of p are disjoint from $\mathcal{B}$ and for every collider v_k on p, there is a node $b^k \in \mathcal{B}$ such that $b^k \in \mathrm{DE}\, v_k$ in $\mathcal{G}_{\text{full}}$.

First, observe, that the end node m is not in M: It is not in $\{Y\}$ (since it was found among the ancestors of Y), and by construction of R, it is not in $\mathcal{B} \cup \mathcal{X} \cup \mathcal{Z}$. Also $m \notin \mathcal{I}$: By the construction of R through recursive rewriting as parents, there exists a directed path from m to Y that does not intersect $\mathcal{X} \cup \mathcal{Z} \cup \mathcal{B}$; if $m \in \mathcal{I}$ this path would violate requirement (CIV1').

Let $w_1, \ldots, w_L$ be those vertices among $v_1, \ldots, v_n$ that appear in the marginalized graph $\mathcal{G}_M$ and let $w_0 \coloneqq i$ and $w_{L+1} \coloneqq Y$. We show that a path including the nodes $w_0, \ldots, w_{L+1}$ (and possibly some additional colliders, see below) from $\mathcal{I}$ to Y exists in $\mathcal{G}_M$; this path is still unblocked, given $\mathcal{B}$, if we remove any outgoing edges from $\mathcal{X} \cup \mathcal{Z}$ that lie on a directed path to Y, creating a contradiction to requirement (CIV1').

For $0 \leq k \leq L$ consider the segment of p (as a path in $\mathcal{G}_{\text{full}}$) from $w_k - \cdots - w_{k+1}$, where $\cdots$ represent edges v_i from p that are not in M. If there are no colliders on this segment, by Definition 1 at least one of the edges $w_k \to w_{k+1}, w_k \leftarrow w_{k+1}$ or $w_k \leftrightarrow w_{k+1}$ are present in $\mathcal{G}_M$. If there is exactly one collider on $w_k - \cdots - w_{k+1}$ (as part of p in $\mathcal{G}_{\text{full}}$), the segment must be one of the following four options:

$$
\begin{aligned}
w_k \to \cdots \to v_i \leftarrow \cdots \leftarrow w_{k+1}, \\
w_k \leftarrow \cdots \leftarrow v_{j_1} \to \cdots \to v_i \leftarrow \cdots \leftarrow w_{k+1}, \\
w_k \to \cdots \to v_i \leftarrow \cdots \leftarrow v_{j_2} \to \cdots \to w_{k+1}, \\
w_k \leftarrow \cdots \leftarrow v_{j_1} \to \cdots \to v_i \leftarrow \cdots \leftarrow v_{j_2} \to \cdots \to w_{k+1},
\end{aligned}
$$

where v_{j_1}, v_{j_2} also are nodes on p. But since $\mathrm{DE}\, v_{i\mathcal{G}_{\text{full}}} \cap \mathcal{B} \neq \emptyset$, this implies that at least one of the following paths are present in $\mathcal{G}_M$:

$$
\begin{aligned}
w_k \to b_{k,1} \leftarrow w_{k+1} \\
w_k \leftrightarrow b_{k,1} \leftarrow w_{k+1} \\
w_k \to b_{k,1} \leftrightarrow w_{k+1} \\
w_k \leftrightarrow b_{k,1} \leftrightarrow w_{k+1},
\end{aligned}
$$

where $b_{k,1} \in \mathcal{B}$ (there is no node a from M on the path from v_i to $b_{k,1}$, because if $a \in \mathcal{X} \cup \mathcal{Z} \cup \{Y\}$, requirement (CIV2) would be violated, and if $a \in \mathcal{I}$, this would constitute another path in $\mathcal{G}_{\text{full}}$ from Y to $\mathcal{I}$ that is unblocked given $\mathcal{B}$, using the same argument as for the original path). Similarly, if there are several colliders on $w_k - \cdots - w_{k+1}$, a path $w_k \to b_{k,1} \leftrightarrow \cdots \leftrightarrow b_{k,L} \leftarrow w_{k+1}$ (or one of the configurations $\to \cdots \leftrightarrow$, $\leftrightarrow \cdots \leftarrow$ or $\leftrightarrow \cdots \leftrightarrow$ as first and last edge) is present in $\mathcal{G}_M$, where $b_{k,1}, \ldots, b_{k,L} \in \mathcal{B}$.

We now construct a path p_M in $\mathcal{G}_M$ that is d-connecting $\mathcal{I}$ and Y, given $\mathcal{B}$: For $k = 0, \ldots, L-1$, paste together the segments (in $\mathcal{G}_M$) from w_k to w_{k+1} including those possible colliders $b_{k,j}$ discussed above. Further, add the edge $w_L \to y$ or $w_L \leftrightarrow y$, depending on the orientation of the edge $w_L - m$ in p. If w_k was a collider on p, it is

also a collider on p_M. Since p was unblocked in $\mathcal{G}_{\text{full}}$, given $\mathcal{B}$, p_M is unblocked in $\mathcal{G}_M$, given $\mathcal{B}$.

We now argue that the path is still unblocked in $\mathcal{G}_M$, given $\mathcal{B}$, if we remove the outgoing edges from $\mathcal{X} \cup \mathcal{Z}$ that are on a directed path to Y. Because $m \notin \mathcal{X} \cup \mathcal{Z} \cup \mathrm{DE}\,\mathcal{X} \cup \mathcal{Z}$, p_M does not contain any edge outgoing of $\mathcal{X} \cup \mathcal{Z}$ on a directed path to Y: Indeed, if for some $w_k \in \mathcal{X} \cup \mathcal{Z}$, p_M contained the segment w_k to w_{k+1}, there would be some $k+2 \leq k' \leq L+1$ such that $w_{k'} \notin \mathrm{DE}\,\mathcal{X} \cup \mathcal{Z}_{\mathcal{G}_{\text{full}}}$ (because otherwise m would be a descendant of $\mathcal{X} \cup \mathcal{Z}$). But this would imply that p_M has a collider, which is a descendant of $\mathcal{X} \cup \mathcal{Z}$ and an ancestor of $\mathcal{B}$ (in $\mathcal{G}_M$), which is not possible by requirement (CIV2). Thus, the path p_M is unblocked given $\mathcal{B}$ in the graph where we remove outgoing edges from $\mathcal{X} \cup \mathcal{Z}$ on direct paths to Y from $\mathcal{G}_M$. This contradicts the assumption that requirement (CIV1') is satisfied in $\mathcal{G}_M$.

A similar argument proves (2), that is, $\mathbb{E}[\mathrm{cov}(\varepsilon, \mathcal{I}|\mathcal{B})] = 0$: Each ε^i was accumulated as a noise variable of an ancestor (in $\mathcal{G}_{\text{full}}$) of Y, $A^i \notin \mathcal{X} \cup \mathcal{Z} \cup \mathcal{B}$; by construction, a directed path from $\mathcal{Z} \cup \mathcal{X}$ through A^i to Y exists in $\mathcal{G}_{\text{full}}$ that does not intersect $\mathcal{X} \cup \mathcal{Z} \cup \mathcal{B}$, except at the first node of this path. Hence, $\mathcal{B}$ does not contain a descendant of A^i in $\mathcal{G}_{\text{full}}$ (because that would imply $\mathcal{B}$ containing a descendant of $\mathcal{X} \cup \mathcal{Z}$ in $\mathcal{G}_M$, violating requirement (CIV2)). Also, A^i is not an ancestor of any node in $\mathcal{I}$ in $\mathcal{G}_{\text{full}}$, because that would imply an unblocked path from $\mathcal{I}$ via A_i to Y in $\mathcal{G}_{\text{full}}$ that does not contain any node in $\mathcal{B}$ and therefore this corresponds to an unblocked path in $\mathcal{G}_M$, too. Using the MA(∞)-representation of $\mathcal{B}$ and $\mathcal{I}$, see Hamilton [1994], ε^i is independent of $(\mathcal{B}, \mathcal{I})$, and so $\mathbb{E}[\mathrm{cov}(\varepsilon^i, \mathcal{I}|\mathcal{B})] = 0$. This concludes the proof of the first part.

Part (ii) follows because if the $(d_{\mathcal{X}} + d_{\mathcal{Z}}) \times d_{\mathcal{I}}$ matrix $\mathbb{E}[\mathrm{cov}(\tilde{\mathcal{X}}, \mathcal{I}|\mathcal{B})]$ has rank $d_{\mathcal{X}} + d_{\mathcal{Z}}$, then if a solution to the moment equation $\mathbb{E}[\mathrm{cov}(Y, \mathcal{I}|\mathcal{B})] = \beta \mathbb{E}[\mathrm{cov}(\tilde{\mathcal{X}}, \mathcal{I}|\mathcal{B})]$ exists, it is unique.

For part (iii), let $\bar{\mathbf{X}} \coloneqq [\mathbf{X}^\top \quad \mathbf{Z}^\top]^\top$. By (5), we have to show that

$$\hat{\mathbb{E}}[r_{\mathbf{Y}} r_{\mathbf{I}}^\top]\; W\; \hat{\mathbb{E}}[r_{\mathbf{I}} r_{\bar{\mathbf{X}}}^\top] \left(\hat{\mathbb{E}}[r_{\bar{\mathbf{X}}} r_{\mathbf{I}}^\top]\; W\; \hat{\mathbb{E}}[r_{\mathbf{I}} r_{\bar{\mathbf{X}}}^\top] \right)^{-1} \xrightarrow{P} \gamma$$

with $\gamma \coloneqq [\beta^\top, \alpha^\top]^\top$. From (2) we have that empirical moments converge in probability to the population moment, and thus, using Slutsky's Theorem, we get that

$$\hat{\mathbb{E}}[r_{\mathbf{Y}} r_{\mathbf{I}}^\top]\; W\; \hat{\mathbb{E}}[r_{\mathbf{I}} r_{\bar{\mathbf{X}}}^\top] \left(\hat{\mathbb{E}}[r_{\bar{\mathbf{X}}} r_{\mathbf{I}}^\top]\; W\; \hat{\mathbb{E}}[r_{\mathbf{I}} r_{\bar{\mathbf{X}}}^\top] \right)^{-1} \xrightarrow{P} \mathbb{E}[r_{Y_t} r_{\mathcal{I}_t}^\top]\; W\; \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top] \left(\mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top]\; W\; \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top] \right)^{-1},$$

where $r_{\bar{\mathcal{X}}} = \bar{\mathcal{X}} - \mathbb{E}[\bar{\mathcal{X}}|\mathcal{B}]$ and similarly for r_Y and $r_{\mathcal{I}}$. We can rewrite $\mathbb{E}[r_{Y_t} r_{\mathcal{I}_t}^\top]$ by adding and subtracting $\gamma r_{\bar{\mathcal{X}}_t}$:

$$\begin{aligned} \mathbb{E}[r_{Y_t} r_{\mathcal{I}_t}^\top] &= \mathbb{E}[(r_{Y_t} - \gamma r_{\bar{\mathcal{X}}_t}) r_{\mathcal{I}_t}^\top] + \gamma \mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top] \\ &= 0 + \gamma \mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top]. \end{aligned}$$

The first term is zero due to the conditional uncorrelation established in (i) and we can

thus conclude that

$$\begin{aligned}\mathbb{E}[r_{Y_t} r_{\mathcal{I}_t}^\top]\ W\ \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top]\Big(\mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top]\ W\ \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top]\Big)^{-1} \\ = \gamma \mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top]\ W\ \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top]\Big(\mathbb{E}[r_{\bar{\mathcal{X}}_t} r_{\mathcal{I}_t}^\top]\ W\ \mathbb{E}[r_{\mathcal{I}_t} r_{\bar{\mathcal{X}}_t}^\top]\Big)^{-1} \\ = \gamma.\end{aligned}$$

□

F.4.6. Proof of Proposition 2

Proposition 2 (Failure of naive IV adaption)**.** *Consider a process* $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ *satisfying Assumption (A2) with* $d_I = d_X = d_H = d_Y = 1$*. If* $\operatorname{cov}(X_{t-1}, I_{t-2}) \neq 0$ *and* $\alpha_{I,I}\alpha_{Y,Y} \neq 1$*, the* $\mathrm{IV}_{X_{t-1}\to Y_t}(I_{t-2})$ *estimator* $\hat{\beta}$ *converges in probability to*

$$(1 - \alpha_{I,I}\alpha_{Y,Y})^{-1}\beta.$$

Consequently, $\hat{\beta}$ *is in general not consistent for the causal effect* β *of* X_{t-1} *on* Y_t*, unless* I *or* Y *do not have any autoregressive structure, that is,* $\alpha_{I,I} = 0$ *or* $\alpha_{Y,Y} = 0$*.*

Proof. Since ε_t^Y and H_{t-1} are both independent of I_{t-2} (for H_{t-1}, this follows from Assumption (A2) and Theorem 1), it follows that

$$\begin{aligned}&\mathbb{E}[Y_t I_{t-2}] = \alpha_{Y,Y}\alpha_{I,I}\mathbb{E}[Y_{t-1}I_{t-3}] + \beta\mathbb{E}[X_{t-1}I_{t-2}] \\ \implies\ &\mathbb{E}[Y_t I_{t-2}] = (1 - \alpha_{Y,Y}\alpha_{I,I})^{-1}\beta\mathbb{E}[X_{t-1}I_{t-2}],\end{aligned}$$

where in the last step we use that by covariance stationarity $\mathbb{E}[Y_{t-1}I_{t-3}] = \mathbb{E}[Y_t I_{t-2}]$ (covariance stationarity follows from Assumption (A1), see Hamilton [1994]). The $\mathrm{IV}_{X_{t-1}\to Y_t}(I_{t-2})$ moment equation is $\mathbb{E}[(Y_t - bX_{t-1})I_{t-2}] = 0$, which has the solution (because $d_I = d_X = d_Y = 1$)

$$b = \frac{\mathbb{E}[I_{t-2}Y_t]}{\mathbb{E}[I_{t-2}X_{t-1}]}.$$

By plugging in the expression for $\mathbb{E}[Y_t I_{t-2}]$ above, we get $b = (1 - \alpha_{Y,Y}\alpha_{I,I})^{-1}\beta$. □

F.4.7. Proof of Theorem 4

Theorem 4 (Identification with conditioning set)**.** *Consider a process* $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ *satisfying Assumption (A2). Let either* $\mathcal{B}_t := \{I_{t-3}\}$ *or* $\mathcal{B}_t := \{I_{t-3}, X_{t-2}, Y_{t-1}\}$*. Then, the following three statements hold. (i) The causal effect* β *of* X_{t-1} *on* Y_t *satisfies the CIV moment condition* $\mathbb{E}[\operatorname{cov}(Y_t - \beta X_{t-1}, I_{t-2}|\mathcal{B}_t)] = 0$*. (ii) Furthermore, if* $\mathbb{E}[\operatorname{cov}(X_{t-1}, I_{t-2}|\mathcal{B}_t)]$ *has rank* d_X*, then* β *is identified by* $\mathrm{CIV}_{X_{t-1}\to Y_t}(I_{t-2}|\mathcal{B}_t)$*. (iii) If, additionally,* $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{I}_t$*, and* $\mathbf{B}_t$ *are observations of* X, Y, I *and* $\mathcal{B}$ *at* T *time points, then* β *can be consistently estimated as* $T \to \infty$ *by* $\mathrm{CIV}_{\mathbf{X}_{t-1}\to\mathbf{Y}_t}(\mathbf{I}_{t-2}|\mathbf{B}_t)$*, that is, the output of Algorithm 1.*

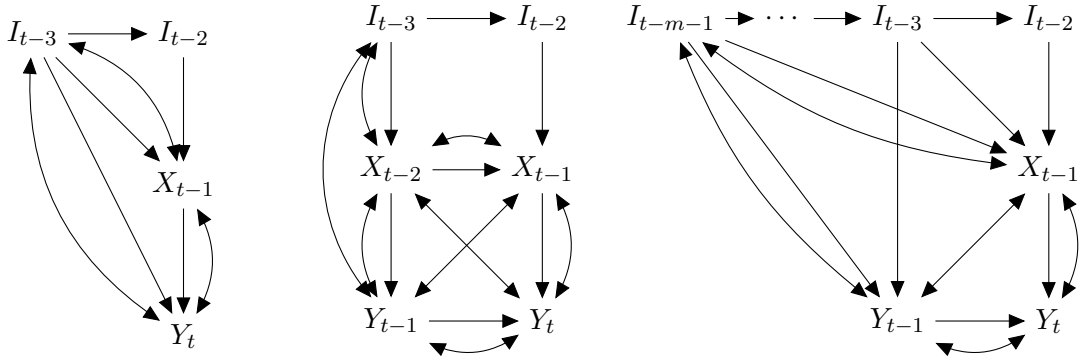


Figure F.3.: (*left*) Marginalization of the full time graph to nodes $I_{t-2}, I_{t-3}, X_{t-1}$ and Y_t. (*middle*) Marginalization of the full time graph to nodes I_{t-2}, X_{t-1} and Y_t and their lagged values. (*right*) Marginalization to m instrument nodes $I_{t-2}, \dots, I_{t-m-1}$, and X_{t-1}, Y_t, and Y_{t-1}.

Proof. By Theorem 3, it suffices for part (i) to show that requirements (CIV1') and (CIV2) are satisfied for $\mathcal{X}_t \coloneqq \{X_{t-1}\}, \mathcal{I}_t \coloneqq \{I_{t-2}\}$, $\mathcal{B}_t$, and Y_t in the marginalized graph $\mathcal{G}_{M_t}$ with $M_t \coloneqq \mathcal{X}_t \cup \mathcal{I}_t \cup \mathcal{B}_t \cup \{Y_t\}$ (with $\mathcal{B}_t$ being either of the two sets from the theorem), see Fig. F.3 *left* and *middle*. For either choice of $\mathcal{B}_t$, $\mathcal{B}_t$ is not a descendant of X_{t-1} and Y_t, so requirement (CIV2) is satisfied (see Fig. F.3 *left* and *middle*). To show that requirement (CIV1') holds, we argue that every path from I_{t-2} to Y_t is blocked by I_{t-3} in $\mathcal{G}_{M_t(X_{t-1} \not\to Y_t)}$, the graph obtained from $\mathcal{G}_{M_t}$ by removing the directed edge from X_{t-1} to Y_t. For either graph, we have that any path from I_{t-2} to Y_t either contains the non-collider I_{t-3} or the collider X_{t-1}. Since I_{t-3} is in the conditioning set $\mathcal{B}_t$ (for either definition of $\mathcal{B}_t$) and $(\{X_{t-1}\} \cup \mathrm{DE}\, X_{t-1}) \cap \mathcal{B}_t = \emptyset$, any path from I_{t-2} to Y_t in $\mathcal{G}_{M_t(X_{t-1} \not\to Y_t)}$ is blocked by $\mathcal{B}_t$.

Parts (ii) and (iii) follow directly from Theorem 3. □

F.4.8. Proof of Theorem 5

Theorem 5 (Identification with nuisance regressor)**.** *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying Assumption (A2). Let $\mathcal{I}_t \coloneqq \{I_{t-2}, \dots, I_{t-m-1}\}$ for an $m \geq 1$ and $\mathcal{Z}_t \coloneqq \{Y_{t-1}\}$. Then, the following three statements hold. (i) There exists $\alpha \in \mathbb{R}$ such that the causal effect β of X_{t-1} on Y_t satisfies the NIV moment condition $\mathbb{E}[\mathrm{cov}(Y_t - \beta X_{t-1} - \alpha\mathcal{Z}_t, \mathcal{I}_t)] = 0$. (ii) Further, if $\mathbb{E}[[X_{t-1}^\top, \mathcal{Z}_t^\top]^\top \mathcal{I}_t^\top]$ has rank $d_X + d_Y$, β is identified by $\mathrm{NIV}_{X_{t-1}\to Y_t}(\mathcal{I}_t, \mathcal{Z}_t)$. (iii) If, additionally, $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{I}_t$, and $\mathbf{Z}_t$ are observations of $X, Y, \mathcal{I}$ and $\mathcal{Z}$ at T time points, then β can be consistently estimated as $T \to \infty$ by $\mathrm{NIV}_{\mathbf{X}_{t-1}\to\mathbf{Y}_t}(\mathbf{I}_t, \mathbf{Z}_t)$, that is, the output of Algorithm 1.*

Proof. By Theorem 3, it suffices for part (i) to show that requirements (CIV1') and (CIV2) are satisfied for $\mathcal{X}_t \coloneqq \{X_{t-1}\}, \mathcal{Z}_t, \mathcal{I}_t$, and Y_t in the marginalized graph $\mathcal{G}_{M_t}$ with $M_t \coloneqq \mathcal{X}_t \cup \mathcal{I}_t \cup \mathcal{Z}_t \cup \{Y_t\}$ (see Fig. F.3 *right*). Since $\mathcal{B} = \emptyset$, requirement (CIV2) is trivially satisfied. It remains to argue that requirement (CIV1) requirement (CIV1') is satisfied, that

is, that $\mathcal{I}_t$ is d-separated from Y_t in the marginalized graph $\mathcal{G}_{M_t(X_{t-1},Y_{t-1}\not\to Y_t)}$ obtained from $\mathcal{G}_{M_t}$ by removing the edges $Y_{t-1} \to Y_t$ and $X_{t-1} \to Y_t$. Let $s \in \{t-m-1, ..., t-2\}$. Every path from I_s to Y_t must go through either the collider $\to X_{t-1} \leftarrow$ or the collider $\to Y_{t-1} \leftarrow$ and since the conditioning set is empty, those paths are blocked.

Parts (ii) and (iii) follow directly from Theorem 3. □

F.4.9. Proof of Theorem 6

A brief review of Jordan canonical forms

If M is an arbitrary square matrix of size $d \times d$, there exists a unique (up to row or column permutations) square invertible matrix Q of the same dimension such that $M = QJQ^{-1}$ where J is a $d \times d$ block diagonal matrix

$$J = J_{m_1}(\lambda_1) \oplus \ldots \oplus J_{m_k}(\lambda_k) := \mathrm{diag}(J_{m_1}(\lambda_1), \ldots, J_{m_k}(\lambda_k)) \tag{F.10}$$

with each *Jordan block* $J_{m_i}(\lambda_i)$ being an $m_i \times m_i$ matrix having one value λ_i on the diagonal and ones on the superdiagonal (and zeros elsewhere): that is, for all $m \in \mathbb{N}_{>0}$,

$$J_m(\lambda) := \begin{pmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}.$$

We sometimes write J_m instead of $J_m(\lambda)$ and call (F.10) the Jordan canonical form. Jordan forms and the involved matrices satisfy the following properties [Horn and Johnson, 1985].

- Let N_m (we simply write N if the dimension is obvious) be the canonical nilpotent matrix of degree m, that is the $m \times m$-matrix with ones in the superdiagonal and zeroes elsewhere. Then $J_m(\lambda) = \lambda 1 + N$ and by the binomial formula $J_m^n = \sum_{i=0}^n \binom{n}{i} \lambda^{n-i} N^i$.
- Every diagonal value of a Jordan block is an eigenvalue of M and for every eigenvalue λ of M, there is at least one Jordan block with diagonal λ. There may however be more than one Jordan block for the same eigenvalue.
- The geometric multiplicity of an eigenvalue λ is the number of corresponding Jordan blocks
- The algebraic multiplicity of an eigenvalue λ is the sum of the sizes m_i of the corresponding Jordan blocks.
- If M is diagonalizable, all Jordan blocks are of size one, which is equivalent to the algebraic and geometric multiplicities being equal.

Some Lemmata for the proof of Theorem 6

We say that a vector v in a d-dimensional vector space is *cyclic of the* $d \times d$ *matrix* J if $v, Jv, \ldots, J^{d-1}v$ constitute a basis for the vector space.

Lemma F.1. *Let* $J = J_{m_1}(\lambda_1) \oplus \ldots \oplus J_{m_k}(\lambda_k)$ *be a block Jordan form over* $\mathbb{C}$ *for a square matrix* J*. If two or more blocks have the same eigenvalue, no vector* $v \in \mathbb{R}^{\sum_{i=1}^k m_i}$ *is cyclic of* J*.*

Proof. It suffices to consider the case $J = J_{m_1}(\lambda) \oplus J_{m_2}(\lambda)$ where without loss of generality $m_1 \geq m_2$. For $J_{m_1}(\lambda) = \lambda 1 + N_{m_1}$ and $J_{m_2}(\lambda) = \lambda 1 + N_{m_2}$ the degree m_1 minimal polynomial $p(x) = (x - \lambda)^{m_1}$ annihilates J such that $p(J) = 0$. Consequently J^{m_1} can be written as a linear combination of $J^0, \ldots, J^{m_1 - 1}$. In particular $J^0 v, \ldots, J^{m_1 + m_2 - 1} v$ cannot be linearly independent. □

Lemma F.2. *Let* $J = J_{m_1}(\lambda_1) \oplus \ldots \oplus J_{m_k}(\lambda_k)$ *be a block Jordan form over* $\mathbb{C}$ *for a square matrix* J*, with each block corresponding to a distinct eigenvalue* λ_i*. Then* $v \in \mathbb{C}^{\sum_{i=1}^k m_i}$ *is a cyclic vector for* J *if and only if for each* $d = 1, \ldots, k$ *the entry* $v_{\sum_{i=1}^d m_i}$ *is non-zero.*

Proof. We first show by contraposition that if v is cyclic for J, the corresponding entries will be non-zero. If it does not hold that for each $d = 1, \ldots, k$ the entry $v_{\sum_{i=1}^d m_i}$ is non-zero, we may, without loss of generality, assume that the last entry of v is zero, such that $v = [u, 0]^\top$ for suitable $u \in \mathbb{C}^{-1 + \sum_{i=1}^k m_i}$. Denote $\lambda = \lambda_k$ the eigenvalue corresponding to the last Jordan block $J_{m_k}(\lambda_k)$; observe that the bottom row of J^n is $[0, \ldots, 0, \lambda^n]$ for any power n, and so the last entry of $J^n v$ is 0 for every n and consequently the matrix $[J^0 v, J^1 v, \ldots, J^{(\sum_{i=1}^k m_i) - 1} v]$ has a 0-row. Consequently, v is not cyclic of J. This shows that if v is cyclic, the entries $v_{\sum_{i=1}^d m_i}$ are non-zero.

Now we show the other implication by induction over k. Assume first that $k = 1$, i.e., $J = J_m(\lambda)$ consists of a single Jordan block. A vector $v = [v_1, \ldots, v_m]^\top$ is cyclic of J if $v_m \neq 0$. Indeed, consider coefficients $a_0, \ldots, a_{m-1}$ such that $0 = \sum_{n=0}^{m-1} a_n J^n v$. Recall that $J^n = \sum_{i=0}^n \binom{n}{i} \lambda^{n-i} N^i$, which implies

$$0 = \sum_{n=0}^{m-1} a_n J^n v = \sum_{n=0}^{m-1} a_n \left(\sum_{i=0}^{n} \binom{n}{i} \lambda^{n-i} N^i \right) v = \sum_{i=0}^{m-1} \left(\sum_{n=i}^{m-1} a_n \binom{n}{i} \lambda^{n-i} \right) N^i v,$$

where in the final equality, we swap the order of summation, using that the pairs (n, i) where $n \in \{0, \ldots, m-1\}$ and $i \in \{0, \ldots, n\}$, are the same as the pairs (n, i) where $i \in \{0, \ldots, m-1\}$ and $n \in \{i, \ldots, m-1\}$. Since $v_m \neq 0$ the collection $N^0 v, \ldots, N^{m-1} v$ are linearly independent: they form an upper-triangular matrix with v_m on the diagonal. This implies that, in particular, the coefficient on $N^{m-1} v$ must be 0. But this coefficient equals $\sum_{n=m-1}^{m-1} a_n \binom{n}{m-1} \lambda^{n-(m-1)} = a_{m-1}$, and so $a_{m-1} = 0$. Substituting this into the coefficient on $N^{m-2} v$, one obtains $a_{m-2} = 0$ and so forth. Therefore, $a_n = 0$ for all n and thus $J^0 v, \ldots, J^{m-1} v$ are linearly independent, so v is cyclic of J if $v_m \neq 0$.

Next assume that the induction hypothesis holds for any matrix with k Jordan blocks $J = J_{m_1}(\lambda_1) \oplus \ldots \oplus J_{m_k}(\lambda_k)$ with distinct eigenvalues for each block and for any vector $v = [v_1, \ldots, v_{\sum_{i=1}^{k} m_i}]^\top$ where for every $d = 1, \ldots, k$: $v_{\sum_{i=1}^{d} m_i} \neq 0$. Now consider the additional Jordan block $D = J_{m_{k+1}}(\lambda_{k+1})$ where $\lambda_{k+1} \neq \lambda_1, \ldots \lambda_k$ and the vector u whose last entry $u_{m_{k+1}}$ is non-zero, and let $\tilde{J} = J \oplus D$, $\tilde{v} = [v^\top, u^\top]^\top$.

Define the polynomial p of degree $\sum_{i=1}^{k} m_i$ by $p(x) = (x - \lambda_1)^{m_1}(x - \lambda_2)^{m_2} \cdots (x - \lambda_k)^{m_k}$. Observe that $p(\lambda_{k+1}) \neq 0$ and so $p(D)$ is an upper triangular matrix with $p(\lambda_{k+1})$ on the diagonal. Hence the last entry of the vector $p(D)u$ is $p(\lambda_{k+1})u_{m_{k+1}}$ which is non-zero, and so $p(D)u$ is cyclic of D (by the initial step of the induction proof). Further observe that p annihilates each of the previous blocks because $J_{m_i}(\lambda_i) = \lambda_i 1 + N$ so $(J_{m_i}(\lambda_i) - \lambda_i 1)^{m_i} = N^{m_i} = 0$. Consequently,

$$\begin{aligned} p(\tilde{J}) &= p(J_{m_1}(\lambda_1)) \oplus \ldots \oplus p(J_{m_k}(\lambda_k)) \oplus p(D) \\ &= 0 \oplus \ldots \oplus 0 \oplus p(D), \end{aligned}$$

so that

$$p(\tilde{J})\tilde{v} = [0^\top, \ldots, 0^\top, (p(D)u)^\top]^\top. \tag{F.11}$$

Now to show $\tilde{v}$ is cyclic of $\tilde{J}$, we take any vector $x \in \mathbb{C}^{\sum_{i=1}^{k} m_i}$ and $y \in \mathbb{C}^{m_{k+1}}$. Our aim is to show that $[x^\top, y^\top]^\top$ is in the span of $\tilde{J}^0\tilde{v}, \ldots, \tilde{J}^{(\sum_{i=1}^{k+1} m_i)-1}\tilde{v}$. Since v is cyclic of J, x can be expressed as a linear combination of $J^0 v, \ldots, J^{\sum_{i=1}^{k} m_i - 1} v$. Taking the same linear combination of $\tilde{J}^0\tilde{v}, \ldots, \tilde{J}^{\sum_{i=1}^{k} m_i - 1}\tilde{v}$ yields the vector $[x^\top, z^\top]^\top$ for some $z \in \mathbb{C}^{m_{k+1}}$. Since $p(D)u$ is cyclic of D, we can write $y - z$ as a linear combination of $D^0 p(D)u, D^1 p(D)u, \ldots, D^{m_{k+1}-1} p(D)u$. It follows from (F.11) that by taking the same linear combination of $\tilde{J}^0 p(\tilde{J})\tilde{v}, \ldots, \tilde{J}^{m_{k+1}-1} p(\tilde{J})\tilde{v}$ one obtains $[0^\top, (y - z)^\top]^\top$. Since p is a polynomial of degree $\sum_{i=1}^{k} m_i$, it follows that both $[x^\top, z^\top]^\top$ and $[0^\top, (y - z)^\top]^\top$ lie in the span of $\tilde{J}^0\tilde{v}, \ldots, \tilde{J}^{\sum_{i=1}^{k+1} m_i - 1}\tilde{v}$, and so does $[x^\top, y^\top]^\top$. Since x and y were arbitrary, the entire space is spanned, completing the induction step. □

Proof of Theorem 6

Theorem 6. *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying Assumption (A2). Assume that $d_I = d_Y = 1$ and let $\mathcal{I}_t := \{I_{t-2}, \ldots, I_{t-m-1}\}$, where $m = d_X + d_Y$. Let A_{XY} and A_I be defined as in (10). The following three statements are equivalent:*

1. $\mathsf{rank}\mathbb{E}[[X_{t-1}^\top, Y_{t-1}^\top]^\top \mathcal{I}_t^\top] = d_X + d_Y$.

2. *The matrix $\left[A_{XY}^0 A_I,\ A_{XY}^1 A_I, \ldots, A_{XY}^{d_X} A_I\right]$ is invertible, where A_{XY}^0 is the identity matrix of size $(d_X + d_Y) \times (d_X + d_Y)$.*

3. *Different Jordan blocks of J have different eigenvalues and for all $q \in \{1, \ldots, k\}$, the coefficient $w_{\sum_{i=1}^{q} m_i}$ is non-zero; here, $J = Q^{-1} A_{XY} Q$ is the Jordan normal*

form[1] *of* A_{XY}*, with* k *Jordan blocks* $J = \mathrm{diag}(J_{m_1}(\lambda_1), \ldots, J_{m_k}(\lambda_k))$*, each with size* m_i *and eigenvalue* λ_i*, and* w *are the coefficients of* A_I *in the basis of the generalized eigenvectors* Q*, that is,* $w = Q^{-1}A_I$*.*

Proof. First observe that

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = A_I I_{t-1} + A_{XY} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \underbrace{\begin{pmatrix} \nu_X \\ \nu_Y \end{pmatrix} H_{t-1} + \varepsilon_t^{X,Y}}_{\text{uncorrelated to I}}$$

and consequently:

$$\mathbb{E}\left[\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_t\right] = \mathbb{E}\left[\left(A_I I_{t-1} + A_{XY} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix}\right) \alpha_{I,I} I_{t-1}\right].$$

From this, we obtain

$$\mathbb{E}\left[\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_t\right] = \underbrace{\mathbb{E}[I_t^2]}_{=:v_I} \alpha_{I,I} \underbrace{(1 - \alpha_{I,I} A_{XY})^{-1}}_{=:B^{-1}} A_I.$$

This expression is justified as B is invertible. (Indeed, if $\alpha_{I,I} = 0$, this is trivial. If $\alpha_{I,I} \neq 0$, since $e_1^\top A_1 = \alpha_{I,I} e_1^\top$, where A_1 is coefficient matrix assumed in Assumption (A2) and $e_1 = [1, 0, \ldots, 0]^\top$ is the first unit vector, $\alpha_{I,I}$ is an eigenvalue of A_1^T and thus of A_1 and in particular, by Assumption (A1), it has absolute value strictly smaller than 1. B is degenerate if and only if $A_{XY} - \frac{1}{\alpha_{I,I}}$ is, but this would imply that $\frac{1}{\alpha_{I,I}}$ would be an eigenvalue of A_{XY}, but since the eigenvalues of A_{XY} are also eigenvalues of A_1 (if $A_{XY} v = \lambda v$, then $A_1[0, v^\top]^\top = \lambda[0, v^\top]^\top$) and belong to the interior of the unit circle, $\frac{1}{\alpha_{I,I}}$ cannot be an eigenvalue of A_{XY}.)

By performing the same expansion for $\mathbb{E}[[X_t^\top, Y_t^\top]^\top I_{t-j}]$ for $j \geq 1$ and plugging in the above, we obtain:

$$\begin{aligned}
\mathbb{E}\left(\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_{t-1}\right) &= A_I v_I + A_{XY} \mathbb{E}\left(\begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} I_{t-1}\right) = v_I B^{-1} A_I \\
\mathbb{E}\left(\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_{t-2}\right) &= v_I \left[A_{XY} B^{-1} + \alpha_{I,I} 1\right] A_I \\
\mathbb{E}\left(\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_{t-3}\right) &= v_I \left[A_{XY}^2 B^{-1} + \alpha_{I,I} A_{XY} + \alpha_{I,I}^2 1\right] A_I
\end{aligned}$$

and in general:

$$\mathbb{E}\left(\begin{pmatrix} X_t \\ Y_t \end{pmatrix} I_{t-1-j}\right) = v_I \left[A_{XY}^j B^{-1} + \sum_{k=0}^{j-1} \alpha_{I,I}^{j-k} A_{XY}^k\right] A_I. \tag{F.12}$$

[1] See Appendix F.4.9 for the definition of Jordan normal forms and the notation that we use.

The columns (denote the j'th column by col_j) of $\Sigma := \mathbb{E}[[X_{t-1}, Y_{t-1}]\mathcal{I}_t^\top]$ are exactly those given by (F.12). If we deduct $\alpha_{I,I}\text{col}_{j-1}$ from col_j we obtain:

$$\begin{aligned}
&v_I \left[\left(\sum_{k=0}^{j-1} \alpha_{I,I}^{j-k} A_{XY}^k + A_{XY}^j B^{-1} \right) - \alpha_{I,I} \left(\sum_{k=0}^{j-2} \alpha_{I,I}^{j-1-k} A_{XY}^k + A_{XY}^{j-1} B^{-1} \right) \right] A_I \\
&= v_I \left[\alpha_{I,I} A_{XY}^{j-1} + A_{XY}^j B^{-1} - \alpha_{I,I} A_{XY}^{j-1} B^{-1} \right] A_I \\
&= v_I (1 - \alpha_{I,I}^2) A_{XY}^j B^{-1} A_I.
\end{aligned}$$

Since deducting columns from each other does not change the determinant, we can create a simpler matrix, Σ_{equiv}, with the same determinant: for $j \in \{2, \ldots, k\}$ we deduct $\alpha_{I,I}\text{col}_{j-1}$ from col_j (starting with the largest j, that is first deducting $\alpha_{I,I}\text{col}_{k-1}$ from col_k, etc.), and obtain

$$\Sigma_{\text{equiv}} = v_I \left[A_{XY}^0 B^{-1} A_I, \quad (1 - \alpha_{I,I}^2) A_{XY}^1 B^{-1} A_I, \quad \ldots, \quad (1 - \alpha_{I,I}^2) A_{XY}^{d_X} B^{-1} A_I \right].$$

By the Laplace expansion, removing $(1-\alpha_{I,I}^2)$ terms appearing in all but the first column scales the determinant by a factor $\frac{1}{(1-\alpha_{I,I}^2)^{d_X}}$, but it will not change its invertibility (from the requirement on the eigenvalues in Assumption (A1), it follows that $1 - \alpha_{I,I}^2 > 0$). The same applies to $v_I = \mathbb{E}[I_t^2]$. Hence, Σ is invertible if and only if

$$\Sigma_{\text{equiv},2} \coloneqq \left[A_{XY}^0 B^{-1} A_I, \quad A_{XY}^1 B^{-1} A_I, \quad \ldots, \quad A_{XY}^{d_X} B^{-1} A_I \right]$$

is invertible. Now observe that B^{-1} commutes with A_{XY}^j. This follows because $BA_{XY} = (1 - \alpha_{I,I} A_{XY}) A_{XY} = A_{XY}(1 - \alpha_{I,I} A_{XY}) = A_{XY} B$. This implies $B^{-1} A_{XY} = A_{XY} B^{-1}$, because for any matrix M where $MB = BM$, it follows that

$$M = MBB^{-1} = BMB^{-1} \implies B^{-1}M = B^{-1}BMB^{-1} = MB^{-1}.$$

This implies that

$$\Sigma_{\text{equiv},2} = B^{-1} \underbrace{\left[A_{XY}^0 A_I, \quad \ldots, \quad A_{XY}^{d_X} A_I \right]}_{=:\Sigma_{\text{equiv},3}}.$$

Since B^{-1} is invertible, it has non-zero determinant, and again invertibility of Σ is equivalent to invertibility of $\Sigma_{\text{equiv},3}$.

Let $A_{XY} = QJQ^{-1}$ be the Jordan block factorization. Observe that

$$\left\{ A_{XY}^0 A_I, \ldots, A_{XY}^{d_X} A_I \right\} = Q \left\{ J^0 Q^{-1} A_I, \ldots, J^{d_X} Q^{-1} A_I \right\}.$$

And finally, since Q is invertible, invertibility of Σ is equivalent to $Q^{-1}A_I$ being cyclic of J. According to Lemma F.1 if two or more Jordan blocks have the same eigenvalue, no vector can be cyclic, so in particular not $Q^{-1}A_I$. If on the contrary no eigenvalue is

shared across Jordan blocks (equivalently, the geometric multiplicity of every eigenvalue is 1), it follow from Lemma F.2 that $Q^{-1}A_I$ is a cyclic vector of J if and only if the vector $Q^{-1}A_I$ is non-zero in the entries indexed by $\sum_{i=1}^{d} m_i$ for all $d = 1, \dots, k$. Writing $A_I = Qa$ in the basis of the columns of Q for some coefficient vector $a \in \mathbb{C}^{d_X+1}$, this means $Q^{-1}A_I$ is a cyclic vector for J exactly when the coefficients $a_{\sum_{i=1}^{d} m_i}$ are non-zero for all $d = 1, \dots, k$. This concludes the proof. □

F.4.10. Proof of Corollary 1

Corollary 1. *Consider a VAR(1) process S with $d_I = 1$ and parameter matrix A, and assume that sparsity pattern of A is given by Assumption (A2) and that the non-zero entries of A are drawn from any distribution which has density with respect to Lebesgue measure. Then β is identifiable with probability* 1.

Proof. We check the conditions of Theorem 6. Since the entries are drawn from a density with respect to Lebesgues measure, the eigenvalues are almost surely distinct. Thus, taking into account the sparsity pattern, A_{XY} can almost surely be diagonalized and the corresponding Jordan form has blocks of size one, all with distinct eigenvalues.

Also with probability one, $w = Q^{-1}\begin{pmatrix}\alpha_{X,I}\\ 0\end{pmatrix}$ does not have any zeroes: Q is determined from A_{XY} (so Q depends only on $\alpha_{X,X}, \alpha_{Y,Y}$ and β), and so the probability that $[\alpha_{X,I}^\top, 0]^\top$ is orthogonal to any of the rows of Q^{-1} is 0. □

F.4.11. Proof of Corollary 2

Corollary 2. *Consider a process S satisfying Assumption (A2) with $d_I > 1$ instrument processes $I^{(1)}, \dots, I^{(d_I)}$. Assume that there is at least one instrument process $I^{(j)}$ such that both of the following conditions hold.*

1. *$I_t^{(j)}$ is independent of $I_s^{(i)}$ for all t, s and $i \neq j$, and*
2. *the requirements of Theorem 6 are satisfied for the reduced process $(I^{(j)}, X, Y)$.*

Then β is identifiable.

Proof. Although the instruments $I^{(i)}, j \neq i$ are observed, we may treat them as latent, being part of the latent process $\tilde{H}_t := (H_t, I_t^{(i,i\neq j)})$. By i), $I^{(j)}$ is independent of $\tilde{H}$. By ii), Theorem 5, and Theorem 6, β is identifiable in the reduced process $(I^{(j)}, X, Y)$, and the solution is therefore also unique in the full system (I, X, Y). □

F.4.12. Proof of Proposition 3

Proposition 3 (Identification with conditioning set relaxing the VAR assumption)**.** *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying (11) and Assumptions (A1') to (A3'). Let $\mathcal{B}_t$ be a set of variables satisfying $\mathrm{PA}(I_{t-2}) \subseteq \mathcal{B}_t \subseteq \mathrm{ND}(Y_t) \cap \mathrm{ND}(I_{t-2})$ in $\mathcal{G}_{full}$. Then, (i), (ii) and (iii) from Theorem 4 hold.*

Proof. To obtain $\mathbb{E}[\mathrm{cov}((Y_t - \beta X_{t-1})I_{t-2}^\top | \mathcal{B}_t)] = 0$, we show that

$$Y_t - \beta X_{t-1} \perp\!\!\!\perp I_{t-2} \,|\, \mathcal{B}_t.$$

Using that $Y_t - \beta X_{t-1} = \alpha_{Y,Y} Y_{t-1} + g(\varepsilon_t^Y, H_{t-1})$ it suffices to show that $Y_{t-1} \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$ and $(\varepsilon_t^Y, H_{t-1}) \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$ since g is measurable. For the first conditional independence we use that $Y_{t-1}, \mathcal{B}_t \subseteq \mathrm{ND}(I_{t-2})$ and $\mathrm{PA}(I_{t-2}) \subseteq \mathcal{B}_t$ to conclude $Y_{t-1} \perp_d I_{t-2} | \mathcal{B}_t$ in $\mathcal{G}_{\text{full}}$ (Indeed, any path from I_{t-2} to Y_{t-1} leaves I_{t-2} either through a parent of I_{t-2} or must contain a collider that is a descendant of I_{t-2}.) By the global Markov property, contained in Assumption (A3'), this implies $Y_{t-1} \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$. For the second conditional independence, we show $\varepsilon_t^Y \perp\!\!\!\perp I_{t-2} | (\mathcal{B}_t, H_{t-1})$ and $H_{t-1} \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$ and use the contraction property of conditional independence to obtain $(\varepsilon_t^Y, H_{t-1}) \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$. Now, $H_{t-1} \perp\!\!\!\perp I_{t-2} | \mathcal{B}_t$ holds by the global Markov property since $H_{t-1} \in \mathrm{ND}(I_{t-2})$ and $\mathrm{PA}(I_{t-2}) \subseteq \mathcal{B}_t$. To show $\varepsilon_t^Y \perp\!\!\!\perp I_{t-2} | (\mathcal{B}_t, H_{t-1})$, we use that by Assumption (A3') ε_t^Y is independent of any finite subset of $\mathrm{ND}(Y_t)$ in $\mathcal{G}_{\text{full}}$. We have that $\mathcal{B}_t \cup \{H_{t-1}\} \cup \{I_{t-2}\} \subseteq \mathrm{ND}(Y_t)$ and thus by weak union we get $\varepsilon_t^Y \perp\!\!\!\perp I_{t-2} | (\mathcal{B}_t, H_{t-1})$ as desired. This proves part (i).

Part (ii) follows because the moment equation $\mathbb{E}[\mathrm{cov}((Y_t - \beta X_{t-1})I_{t-2}^\top | \mathcal{B}_t)] = 0$ is the same as in Theorem 4, and so the rank requirement for identifiability is also the same.

To show part (iii), we have to show that

$$\hat{\mathbb{E}}[r_{\mathbf{Y}_t} r_{\mathbf{I}_{t-2}}^\top] \; W \; \hat{\mathbb{E}}[r_{\mathbf{I}_{t-2}} r_{\mathbf{X}_{t-1}}^\top] \left(\hat{\mathbb{E}}[r_{\mathbf{X}_{t-1}} r_{\mathbf{I}_{t-2}}^\top] \; W \; \hat{\mathbb{E}}[r_{\mathbf{I}_{t-2}} r_{\mathbf{X}_{t-1}}^\top] \right)^{-1} \xrightarrow{P} \beta.$$

This is analogous to the argument in the proof of Theorem 3, except for that the convergence of empirical moments is now guaranteed by Assumption (A2') (instead of Assumption (A1)). Hence, using Slutsky's Theorem and rewriting Y_t as in Theorem 3, gives the desired convergence. □

F.4.13. Proof of Proposition 4

Proposition 4 (Identification with nuisance regressor relaxing the VAR assumption)**.** *Consider a process $S = [I_t^\top, X_t^\top, H_t^\top, Y_t^\top]_{t\in\mathbb{Z}}^\top$ satisfying (11) and Assumptions (A1'), (A2') and (A4'). Let $\mathcal{Z}_t \coloneqq \{Y_{t-1}\}$ and $\mathcal{I}_t \coloneqq \{I_{t-2}, \dots, I_{t-m-1}\}$ for an $m \geq 1$. Then, (i), (ii), and (iii) from Theorem 5 hold.*

Proof. By (11), we have that $Y_t - \beta X_{t-1} - \alpha_{Y,Y} Y_{t-1} = g(\varepsilon_t^Y, H_{t-1})$. Furthermore, by Assumption (A4') $(\varepsilon_t^Y, H_{t-1}) \perp\!\!\!\perp \mathcal{I}_t$. Combining this (and using measureability of g) we obtain

$$Y_t - \beta X_{t-1} - \alpha_{Y,Y} Y_{t-1} \perp\!\!\!\perp \mathcal{I}_t.$$

Thus $\mathbb{E}[(Y_t - \beta X_{t-1} - \alpha_{Y,Y} Y_{t-1})\mathcal{I}_t^\top] = 0$ for $\alpha = \alpha_{Y,Y}$, and part (i) follows.

Part (ii) follows because the moment equation $\mathbb{E}[(Y_t - \beta X_{t-1} - \alpha_{Y,Y} Y_{t-1})\mathcal{I}_t^\top] = 0$ is the same as in Theorem 5, and so the rank requirement for identifiability is also the same.

To show part (iii), let $\bar{\mathbf{X}}_{t-1} := [{\mathbf{X}_{t-1}}^\top, \ {\mathbf{Y}_{t-1}}^\top]^\top$. We have to show that

$$\hat{\mathbb{E}}[\mathbf{Y}_t \boldsymbol{\mathcal{I}}_t^\top] \ W \ \hat{\mathbb{E}}[\boldsymbol{\mathcal{I}}_t \bar{\mathbf{X}}_{t-1}^\top] \Big(\hat{\mathbb{E}}[\bar{\mathbf{X}}_{t-1} \boldsymbol{\mathcal{I}}_t^\top] \ W \ \hat{\mathbb{E}}[\boldsymbol{\mathcal{I}}_t \bar{\mathbf{X}}_{t-1}^\top] \Big)^{-1} \xrightarrow{\mathcal{P}} \gamma$$

with $\gamma := [\beta, \ \ \alpha_{Y,Y}]$. Assumption (A2') ensures convergence of the empirical moments to population moments, and using Slutsky's Theorem in combination with the expression for Y_t, the statement follows as in the proof of Theorem 3. □

F.4.14. Proof of Proposition 5

Proposition 5. *Consider a process $S = [S_t]_{t \in \mathbb{Z}}$ satisfying Assumption (A2). Let β be the causal effect from X_t to Y_{t+1}, and let for an arbitrary $m, \ell \in \mathbb{N}$ $(\alpha_{Y,X}, \alpha_{Y,Y})$ be the population vector of coefficients when regressing $Y_{s+1} - \beta X_s$ on $\{X_{s-k}, k = 1, \dots, m\} \cup \{Y_{s-j}, j = 0, \dots, l\}$. Then*

$$(\alpha_{Y,Y}, \beta, \alpha_{Y,X}) = \arg\min_{a,b,c} \mathbb{E}_{\mathrm{do}(X_t := x)} \left\{ Y_{t+1} - \sum_{j=0}^{\ell} a_j Y_{t-j} - b X_t - \sum_{k=1}^{m} c_k X_{t-k} \right\}^2 .$$

Proof. Recall that β and $\alpha_{Y,Y}, \alpha_{Y,H}$ denote the causal effects from X_t, Y_t and H_t, re-

spectively, to Y_{t+1}. We have

$$
\begin{aligned}
&\min_{a,b,c} \mathbb{E}_{do(X_t:=x)} \left(Y_{t+1} - \sum_{j=0}^{\ell} a_j Y_{t-j} - bX_t - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 \\
&= \min_{a,b,c} \mathbb{E}_{do(X_t:=x)} \left(\{\beta X_t + \alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y\} - \sum_{j=0}^{\ell} a_j Y_{t-j} - bX_t - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 \\
&= \min_{a,b,c} \mathbb{E}_{do(X_t:=x)} (\beta X_t - bX_t)^2 \\
&\qquad + \mathbb{E}_{do(X_t:=x)} (\beta X_t - bX_t) \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right) \\
&\qquad + \mathbb{E}_{do(X_t:=x)} \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 \\
&= \min_{a,b,c} \mathbb{E}_{do(X_t:=x)} (\beta X_t - bX_t)^2 \\
&\qquad + \mathbb{E}_{do(X_t:=x)} (\beta X_t - bX_t) \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right) \\
&\qquad + \mathbb{E} \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 \\
&= \min_{a,b,c} \mathbb{E}_{do(X_t:=x)} (\beta X_t - bX_t)^2 + (\beta x - bx) \mathbb{E} \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right) \\
&\qquad + \mathbb{E} \left(\alpha_{Y,Y} Y_t + \alpha_{Y,H} H_t + \varepsilon_{t+1}^Y - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 \\
&= \min_{a,c} \mathbb{E} \left(Y_{t+1} - \beta X_t - \sum_{j=0}^{\ell} a_j Y_{t-j} - \sum_{k=1}^{m} c_k X_{t-k} \right)^2 .
\end{aligned}
$$

Here, the third and fourth equality signs hold because the joint distribution of the variables H_t, ε_{t+1}^Y, $Y_t, \ldots, Y_{t-\ell}$, and $X_{t-1}, \ldots, X_{t-m}$ is the same under the observational and the intervention distribution – as the variables are all non-descendants of X_t. Further, the minimum is obtained for $b = \beta$ and a and c being the coefficients after (linearly) projecting $Y_{t+1} - \beta X_t$ on the space spanned by $Y_t, \ldots, Y_{t-\ell}$ and $X_{t-1}, \ldots, X_{t-m}$. □

Bibliography

S. Acid and L. M. De Campos. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20 (1):46–63, 1949.

J. D. Angrist and G. W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

J. D. Angrist and A. B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021. Survey Track.

I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. Palmer, B. J. Price, S. Purkayastha, A. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, H. Zhang, and J. W. Gichoya. Reading race: AI recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, July 2021.

E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in Neural Information Processing Systems*, 27, 2014.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

A. Bellot and M. van der Schaar. Accounting for unobserved confounding in domain generalization. *arXiv (2007.10653)*, July 2020.

T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

D. Bertsimas and C. McCord. Optimization over continuous and multi-dimensional decisions with observational data. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138, 2009.

R. Bhattacharya. On actually testing generalized independence (Verma) constraints, 2019. https://www.cs.jhu.edu/~rohit/posts/verma.pdf, last accessed 15.04.2021.

R. Bhattacharya and R. Nabi. On testability of the front-door model via Verma constraints. In *Proceedings of the 38th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

P. J. Bickel, F. Götze, and W. R. van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. In *Selected works of Willem van Zwet*, pages 267–297. Springer, 2012.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013.

J. Bound, C. Brown, and N. Mathiowetz. Chapter 59: Measurement Error In Survey Data. *Handbook of Econometrics*, 5:3705–3843, 2001.

R. J. Bowden and D. A. Turkington. *Instrumental Variables*, volume 8 of *Econometric Society Monographs*. Cambridge University Press, 1985.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

C. Brito and J. Pearl. Generalized instrumental variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 85–93, 2002a.

C. Brito and J. Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002b.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 2nd edition, 1991.

P. Bühlmann and D. Ćevid. Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88(S1):S114–S134, 2020.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

A. Chesher. Identification in nonseparable models. *Econometrica*, 71(5):1405–1441, 2003.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (accepted)*, 2021.

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.

A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.

I. W. P. Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

J. Correa and E. Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020.

R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, pages 115–137, 2003.

D. Danks and S. Plis. Learning causal structure from undersampled time series, 2013. URL https://www.andrew.cmu.edu/user/ddanks/papers/DanksPlis-Final.pdf. Results were presented at NIPS 2013 workshop on causality; last visit of website: 02.03.2022.

T. R. Dawber, G. F. Meadors, and F. E. Moore Jr. Epidemiological approaches to heart disease: the Framingham study. *American Journal of Public Health and the Nations Health*, 41(3):279–286, 1951.

A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

A. A. de Kroon, D. Belgrave, and J. M. Mooij. Causal discovery for causal bandits utilizing separating sets. *arXiv preprint arXiv:2009.07916*, 2020.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 245–264, 2008.

V. Didelez, S. Meng, and N. A. Sheehan. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40, 2010.

M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348, 2009.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

J. C. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv (2007.13982)*, pages 1–39, 2020.

M. Dudik, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104. ACM, 2011.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

R. C. Fair. The estimation of simultaneous equation models with lagged endogenous variables and first order serially correlated errors. *Econometrica: Journal of the Econometric Society*, pages 507–516, 1970.

H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

T. Fernández, A. Gretton, D. Rindt, and D. Sejdinovic. A kernel log-rank test of independence for right-censored data. *Journal of the American Statistical Association*, 0 (0):1–12, 2021.

S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and S. Saria. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 385(3):283–286, July 2021.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

C. Frost and S. G. Thompson. Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A*, 163(2):173–189, 2000.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2008.

W. A. Fuller. *Measurement Error Models*. John Wiley and Sons Inc., 1987.

J. L. Gamella and C. Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

S. Gaspers and S. Mackenzie. On the number of minimal separators in graphs. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 116–121. Springer, 2015.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.

C. W. J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(1):329–352, 1980.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

C. Guestrin, D. Koller, and R. Parr. Multiagent planning with factored MDPs. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *J. Artif. Int. Res.*, 19(1):399–468, 2003.

R. Guo, P. Zhang, H. Liu, and E. Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv (2101.07732)*, January 2021.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

A. R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1st edition, 1994.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

B. Hao, X. Ji, Y. Duan, H. Lu, C. Szepesvari, and M. Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4074–4084. PMLR, 2021.

J. Hartung. A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(7):849–855, 1999.

T. J. Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):1–35, 2018.

L. Henckel. *Graphical Tools for Efficient Causal Effect Estimation*. PhD thesis, ETH Zurich, Zurich, 2021.

M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

M. A. Hernán and J. M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

M. A. Hernán and J. M. Robins. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC, 2020.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, 2009.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

D. R. Hyslop and G. W. Imbens. Bias from classical and other forms of measurement error. *Journal of Business and Economic Statistics*, 19(4):475–481, 2001.

A. Hyttinen, S. Plis, M. Järvisalo, F. Eberhardt, and D. Danks. Causal discovery from subsampled time series data by constraint optimization. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models (PGM)*, pages 216–227, 2016.

G. W. Imbens and W. K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, New York, NY, 2015.

M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE. *The Econometrics Journal*, 25(2):404–432, 10 2021.

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. PMLR, 03–06 Aug 2020.

A. Jonsson and A. Barto. Causal graph based decomposition of factored MDPs. *Journal of Machine Learning Research*, 7(81):2259–2301, 2006.

M. I. Jordan. Artificial intelligence — the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1), 2019.

N. Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2018.

N. Kallus and A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020.

P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does Invariant Risk Minimization Capture Invariance? *arXiv (2101.01134)*, 2021. URL http://arxiv.org/abs/2101.01134.

M. Kearns and D. Koller. Efficient reinforcement learning in factored MDPs. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, page 740–747. Morgan Kaufmann Publishers Inc., 1999.

P. Kemmeren, K. Sameith, L. A. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

L. Kook, B. Sick, and P. Bühlmann. Distributional anchor regression. *Statistics and Computing*, 32(3):1–19, 2022.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). *arXiv (2003.00688)*, March 2020.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

J. A. Laurie, C. G. Moertel, T. R. Fleming, H. S. Wieand, J. E. Leigh, J. Rubin, G. W. McCormack, J. B. Gerstner, J. E. Krook, and J. Malliard. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, 7(10):1447–1456, 1989.

S. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

S. L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, 1996.

S. Lee and E. Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

S. Lee, J. D. Correa, and E. Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv e-prints (2005.01643)*, 2020.

X. Liang, S. Li, S. Zhang, H. Huang, and S. X. Chen. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121, 2016.

Y. Liu and J. Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

M. C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

M. C. Lovell. A simple proof of the FWL theorem. *Journal of Economic Education*, 39 (1):88–91, 2008.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2005.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A. R. Mahmood, H. P. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

G. G. Martinet, A. Strzalkowski, and B. Engelhardt. Variance minimization in the wasserstein space for invariant causal prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 8803–8851. PMLR, 2022.

S. McGrath, J. G. Young, and M. A. Hernán. Revisiting the g-null paradox. *Epidemiology*, 33(1):114–120, 2022.

C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, 1995.

N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

W. Miao and E. T. Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *arXiv (1808.04945)*, 2018.

C. G. Moertel. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: A final report. *Annals of Internal Medicine*, 122(5):321, 1995.

P. Mogensen, N. Thams, and J. Peters. Invariant ancestry search. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15832–15857. PMLR, 2022.

S. W. Mogensen and N. R. Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

A. I. Naimi, E. E. Moodie, N. Auger, and J. S. Kaufman. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*, pages 292–299, 2014.

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation. *Econometrica*, 55(3):703–708, 1987.

C. Nowzohour, M. H. Maathuis, R. J. Evans, and P. Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017.

M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets under covariate shift. *ArXiv e-prints (2106.09848)*, 2021.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. The Morgan Kaufmann series in representation and learning. Morgan Kaufmann, 2014.

J. Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 154–162, 2013.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

J. Peters, S. Bauer, and N. Pfister. Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl (to appear, ArXiv e-prints (2001.06208))*. ACM, 2022.

N. Pfister, S. Bauer, and J. Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019a.

N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019b.

N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann. Stabilizing variable selection and regression. *Annals of Applied Statistics (accepted)*, 2021.

I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM review*, 49(3):371–418, 2007.

M. J. Powell. The NEWUOA software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer, 2006.

D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 417–424, 2001.

J. Quionero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer. *Dataset Shift in Machine Learning*. MIT Press, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

O. Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.

A. Reisach, C. Seiler, and S. Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical report, Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper, 2013.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *ArXiv e-prints (1701.06686)*, 2017.

X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, 2011.

J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

J. M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

J. M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 349–405. MIT Press, 1999.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and L. Wasserman. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, page 409–420. Morgan Kaufmann Publishers Inc., 1997.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

E. Rosenfeld and A. Risteski. The Risks of Invariant Risk Minimization. *arXiv (2010.05761)*, 2020.

T. J. Rothenberg. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591, 1971.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

D. B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398): 543–546, 1987.

L. Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

B. Rüger. Das maximale signifikanzniveau des tests. *Metrika*, 25:171–178, 1978.

S. Saengkyongam, N. Thams, J. Peters, and N. Pfister. Invariant policy learning: A causal perspective. *arXiv preprint arXiv:2106.00808*, 2021.

S. Saengkyongam, L. Henckel, N. Pfister, and J. Peters. Exploiting independent instruments: Identification and distribution generalization. *arXiv preprint arXiv:2202.01864*, 2022.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

M. Schlegel, W. Chung, D. Graves, J. Qian, and M. White. Importance resampling for off-policy prediction. *arXiv preprint arXiv:1906.04328*, 2019.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

C. Schultheiss, P. Bühlmann, and M. Yuan. Higher-order least squares: assessing partial goodness of fit of linear regression. *arXiv preprint arXiv:2109.14544*, 2021.

R. Sen, K. Shanmugam, M. Kocaoglu, A. Dimakis, and S. Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY [u.a.], 1980. ISBN 0471024031.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

C. Shi, T. Xu, W. Bergsma, and L. Li. Double generative adversarial networks for conditional independence testing. *J. Mach. Learn. Res.*, 22:285–1, 2021.

X. Shi, W. Miao, J. C. Nelson, and E. J. T. Tchetgen. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *arXiv (1808.04906)*, 2018.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758.

I. Shpitser and J. Pearl. Dormant independence. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1081–1087. AAAI Press, 2008.

I. Shpitser, T. Richardson, and J. Robins. Testing edges by truncations. In *IJCAI-09 - Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1957–1963, January 2009.

I. Shpitser, T. S. Richardson, J. M. Robins, and R. Evans. Parameter and structure learning in nested Markov models. *ArXiv e-prints (1207.5058)*, 2012.

I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41:3–39, 2014.

Ø. Skare, E. Bølviken, and L. Holden. Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737, 2003.

A. F. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

A. Sonar, V. Pacelli, and A. Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*, pages 21–33. PMLR, 2021.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

C. Squires, Y. Wang, and C. Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

M. Srivastava, T. Hashimoto, and P. Liang. Robustness to Spurious Correlations via Human Annotations. *37th International Conference on Machine Learning*, 2020.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica: journal of the Econometric Society*, pages 557–586, 1997.

A. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

A. Subbaswamy, R. Adams, and S. Saria. Evaluating model robustness and stability to dataset shift. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 13–15 Apr 2021.

A. Subbaswamy, B. Chen, and S. Saria. The stability and accuracy tradeoff under dataset shift: A causal graphical analysis. *Journal of Causal Inference*, to appear, 2022.

M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation.* MIT Press, 2012.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

R. Sutton and A. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 814–823. PMLR, 2015a.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3231–3239. Curran Associates, Inc., 2015b.

K. Takata. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158:1660–1667, 2010.

E. J. Tchetgen Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An Introduction to Proximal Causal Learning. *arXiv (2009.10982)*, 2020.

G. Tennenholtz, U. Shalit, and S. Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

G. Tennenholtz, U. Shalit, S. Mannor, and Y. Efroni. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pages 430–439. PMLR, 2021.

J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

N. Thams and N. R. Hansen. Local independence testing for point processes. *arXiv preprint arXiv:2110.12709*, 2021.

N. Thams, S. Saengkyongam, N. Pfister, and J. Peters. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

N. Thams, M. Oberst, and D. Sontag. Evaluating robustness to dataset shift via parametric robustness sets. In *Neural Information Processing Systems (NeurIPS)*, 2022a. NT and MO contributed equally, order determined by coin flip.

N. Thams, R. Søndergaard, S. Weichwald, and J. Peters. Identifying causal effects using instrumental time series: Nuisance IV and correcting for the past. *arXiv preprint arXiv:2203.06056*, 2022b.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32th International Conference on Machine Learning (ICML)*, pages 2380–2388. PMLR, 2015.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth national conference on Artificial intelligence*, pages 567–573, 2002.

J. Tian, A. Paz, and J. Pearl. Finding minimal d-separators. Technical report, University of California, Los Angeles, 1998.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

R. J. Tibshirani, R. F. Barber, E. J. Candes, and A. Ramdas. Conformal prediction under covariate shift. *ArXiv e-prints (1904.06019)*, 2019.

B. van der Zander, M. Liśkiewicz, and J. Textor. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270: 1–40, 2019.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.

T. S. Verma. Invariant properties of causal models. Technical report, UCLA Cognitive Systems Laboratory, 1991.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

V. Vovk and R. Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer Verlag, Berlin, Heidelberg, 2005.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pages 27–36. PMLR, 2020.

A. S. Weigend. *Time series prediction: Forecasting the future and understanding the past*. Routledge, 2018.

N. Wiener. The theory of prediction. In *Modern Mathematics for Engineers*. McGraw-Hill, 1956.

F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

P. G. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, 1928.

S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5 (3):161–215, 1934.

C. Xie, H. Ye, F. Chen, Y. Liu, R. Sun, and Z. Li. Risk variance penalization. *arXiv (2006.07544)*, June 2020.

A. Yabe, D. Hatano, H. Sumita, S. Ito, N. Kakimura, T. Fukunaga, and K.-i. Kawarabayashi. Causal bandits with propagating inference. In *Proceedings of the*

35th International Conference on Machine Learning, volume 80, pages 5512–5520. PMLR, 2018.

H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

H. Zenati, A. Bietti, M. Martin, E. Diemert, and J. Mairal. Counterfactual learning of continuous stochastic policies. *arXiv preprint arXiv:2004.11722*, 2020.

A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pages 11214–11224. PMLR, 2020.

F. Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813. AUAI Press, 2011.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning*, page 116, 2004.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.