Martin Emil Jakobsen

Causality and Generalizability

Identifiability and Learning Methods

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 2021

Martin Emil Jakobsen
m.jakobsen@math.ku.dk
martin.emil.jakobsen@gmail.com
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Causality and Generalizability: Identifiability and Learning Methods
Supervisor:	Professor Jonas Peters University of Copenhagen
Assessment committee:	Associate Professor Trine Krogh Boomsma (chair) University of Copenhagen
	Professor Søren Hauberg Technical University of Denmark
	Professor Joris Mooij University of Amsterdam
Date of Submission:	August 31, 2021
Date of Defense:	November 4, 2021
ISBN:	978-87-7125-048-0

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen. It was supported by the Carlsberg Foundation.

Abstract

This Ph.D. thesis contains several contributions to the field of statistical causal modeling. Statistical causal models are statistical models embedded with causal assumptions that allow for the inference and reasoning about the behavior of stochastic systems affected by external manipulation (interventions). This thesis contributes to the research areas concerning the estimation of causal effects, causal structure learning, and distributionally robust (out-of-distribution generalizing) prediction methods. We present novel and consistent linear and non-linear causal effects estimators in instrumental variable settings that employ data-dependent mean squared prediction error regularization. Our proposed estimators show, in certain settings, mean squared error improvements compared to both canonical and state-of-the-art estimators. We show that recent research on distributionally robust prediction methods has connections to well-studied estimators from econometrics. This connection leads us to prove that general K-class estimators possess distributional robustness properties. We, furthermore, propose a general framework for distributional robustness with respect to intervention-induced distributions. In this framework, we derive sufficient conditions for the identifiability of distributionally robust prediction methods and present impossibility results that show the necessity of several of these conditions. We present a new structure learning method applicable in additive noise models with directed trees as causal graphs. We prove consistency in a vanishing identifiability setup and provide a method for testing substructure hypotheses with asymptotic family-wise error control that remains valid post-selection. Finally, we present heuristic ideas for learning summary graphs of nonlinear time-series models.

Resumé

Denne Ph.D. afhandling indeholder flere bidrag til forskningsområdet for statistisk kausal modellering. Statistiske kausale modeller er statistiske modeller med kausale antagelser, som muliggør inferens og ræsonnement omkring stokastiske systemers adfærd under ekstern manipulation. Denne afhandling bidrager til forskningsområderne vedrørende estimering af kausale effekter, kausale strukturer og fordelingsrobuste prædiktionsmetoder. Vi præsenterer nye estimatorer for lineære og ikke-lineære kausale effekter i modeller med instrumentelle variabler. Disse estimatorer anvender dataafhængig regulering og viser forbedret gennemsnitlig kvadratfejl sammenlignet med anerkendte metoder. Vi viser, at nyere forskning, om fordelingsrobuste forudsigelsesmetoder har forbindelser til velkendte estimatorer fra økonometri. Vi beviser, at generelle K-klasse estimatorer besidder fordelingsrobuste prædiktions egenskaber. Vi foreslår endvidere en kausal tilgang til fordelingsrobuste prædiktionsmetoder. Vi udleder tilstrækkelige betingelser for identificering af fordelingsrobuste prædiktionsmetoder og viser endvidere nødvendigheden af flere af disse betingelser. Vi præsenterer en ny metode til at estimere kausale strukturer, der kan anvendes i modeller med additiv støj og orienterede træer som kausale grafer. Vi beviser, at metoden er konsistent, og fremstiller metoder til at teste hypoteser omkring den kausale struktur. Endelig præsenterer vi heuristiske ideer til at lære opsummeringsgrafer for ikke-lineære tidsseriemodeller.

Preface

This thesis has been submitted in partial fulfillment of the requirements for the Ph.D. degree at the Department of Mathematical Sciences, Faculty of Science, University of Copenhagen. This work was written between August 2018 and August 2021 at the Copenhagen Causality Lab, Section for Statistics and Probability Theory. This research was funded by The Carlsberg Foundation. While the pandemic threw a wrench in the planned research visit abroad and resulted in approximately half of this work being written within the confines of my apartment, it has nonetheless been a great experience.

Acknowledgments

First and foremost, I would like to thank my supervisor Jonas Peters. It has truly been a pleasure working under your excellent guidance. Your commitment to our projects and our frequent meetings have been invaluable. You always dropped whatever you had in your hands in order to consider and answer my countless questions, no matter how trivial or uninteresting they may have been.

To all my co-authors, Peter Bühlmann, Rune Christiansen, Nicola Gnecco, Phillip Mogensen, Jonas Peters, Lasse Petersen, Niklas Pfister, Rajen Shah, Nikolaj Thams, Gherardo Varando, and Sebastian Weichwald, I thank you for the fruitful collaborations, exciting discussions, and uplifting company. To all my colleagues at the department, thank you for making my time at the department enjoyable. I thank Steffen Lauritzen for helpful discussions about various mathematical problems. I thank all my teachers, in particular Ernst Hansen, Thomas Mikosch and Anders Rønn-Nielsen, who taught me the foundations on which this thesis is written. To my friend Mads Raad, thank you for almost ten years of mathematical null-set discussions.

To all of my friends, who time and time again have been told that I was too busy to hang out, thank you for never stopping to care. I thank my friends and family for their encouraging words and for always taking an interest in my work. To my mother and father, who inspired me to push my boundaries and helped me realize the fun and beauty of mathematics, thank you for your unconditional support and love.

Martin Emil Jakobsen
August, 2021

The thesis has been edited and minor typographical errors has been corrected in agreement with the official guidelines prior to printing. A version of this thesis containing additional corrections can be found at `arxiv.org/abs/2110.01430`.

Martin Emil Jakobsen
October, 2021

Summary of Contributions

This thesis consists of one introductory and four main chapters. The main chapters aim to advance various areas of research within the field of statistical causal modeling. Chapter 1 contains a general introduction to causal modeling and reasoning in the mathematical framework of statistical causal models. We, furthermore, introduce the research topics of later chapters and discuss and summarize our contributions in more detail. The main chapters and their corresponding appendices consist (up to minor corrections and aesthetic modifications) of previously published, forthcoming, or submitted papers. The four main chapters correspond to the following papers:

Chapter 2: Jakobsen, M. E. and Peters, J. Distributional Robustness of K-class Estimators and the PULSE. *The Econometrics Journal (forthcoming)*, 2021. DOI: 10.1093/ectj/utab031.

Chapter 3: Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (forthcoming)*, 2021. DOI: 10.1109/tpami.2021.3094760.

Chapter 4: Jakobsen, M. E., Shah, R., Bühlmann, P., and Peters, J. Structure Learning for Directed Trees. *arXiv preprint arXiv:2108.08871*, 2021.

Chapter 5: Weichwald, S., Jakobsen, M. E., Mogensen, P. B., Petersen, L., Thams, N., and Varando, G. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In Escalante, H. J. and Hadsell, R., editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36. PMLR, 08–14 Dec 2020.

Chapter 2 proposes a novel estimator, called the p-uncorrelated least squares estimator (PULSE), for linear causal effects in instrumental variable (IV) setups. The PULSE can be viewed as a data-dependent mean squared prediction error regularization of the two-stage least squares estimator. We prove that the estimator is consistent, and through simulations studies, we show that in, e.g., weak instrument settings, it is MSE superior to other competing IV causal effect estimators. Furthermore, we establish a connection between K-class estimators from econometrics and the recently proposed anchor regression estimators from the field of out-of-distribution generalizing prediction methods. Prediction methods are

said to be distributionally robust (or out-of-distribution generalizing) with respect to a class of test distributions if it minimizes the worst-case risk over said class. We show that K-class estimators are distributionally robust prediction methods with respect to bounded interventions on exogenous system variables.

In Chapter 3, we propose a general framework for analyzing distributional robustness with respect to test distributions generated by interventions. We provide sufficient conditions for out-of-distribution generalization and present several impossibility results showing the necessity of certain conditions. We propose a nonlinear instrumental variable estimator that uses the previously mentioned data-dependent mean squared prediction error regularization. A simulation study shows that it, in specific setups, is MSE superior to various state-of-the-art nonparametric instrumental variable estimators.

In Chapter 4, we contribute to the field of causal structure learning. We propose a method for learning the causal structure of systems with directed trees as causal graphs. We strengthen established identifiability results of causal graphs for restricted structural causal models. Furthermore, we provide an alternative analysis that proves that for Gaussian noise models, the identifiability of the causal graph is a purely local property of the underlying model. Our learning method does not require heuristic optimization algorithms to recover the causal graph, something that plagues virtually all structure learning methods that do not search for Markov equivalent structures. Furthermore, we prove consistency in an asymptotic setup with decreasing identifiability. We propose a method for testing causal substructure hypotheses. The proposed method has asymptotic family-wise error rate control that remains valid post-selection.

Chapter 5 presents the approaches for learning summary graphs of time-series that won the NeurIPS Causality 4 Climate competition. We articulate our heuristic learning approaches and discuss artifacts of simulated DAG models.

Contents

Chapter 1

Introduction

In many applications, we are interested in reasoning about the behavior of a stochastic system that is affected by external manipulation. For example, in a prediction setup, we may anticipate future external manipulation of the system of interest, such that differences emerge between training and test distributions. Alternatively, we may be interested in the expected changes to a system when we intervene (apply external manipulation) on a system variable. Statistical and probabilistic models are insufficient for such purposes, as they do not possess the formal language and tools to quantify such changes. For such purposes, we need to consider statistical causal models. These are statistical models embedded with causal assumptions that allow us to model and reason about how external manipulation affects the behavior of stochastic systems.

This chapter serves as an introduction to causal modeling and inference. We discuss certain fundamental causal concepts and problems, which hopefully will ease the reading of later chapters for the causally uninitiated reader. We summarize the contributions of the later chapters and explain how they fit within established research in the statistical causal literature.

In Section 1.1, we discuss the difference between the statistical and causal models and introduce some graph terminology used in later chapters. Furthermore, we define structural causal models and introduce the concept of interventions in connection with the assumption of autonomy. In Section 1.2, we discuss the general difficulties with causal inference and explain the necessity of unfalsifiable causal assumptions when inferring causal quantities from observational data. Section 1.3 introduces the independence-based (also called constraint-based) and score-based approaches to causal structure learning and discusses the causal assumptions these approaches need. Section 1.4 introduces the concept of causal effects. Here, we discuss how sufficient knowledge of the underlying causal structure enables the inference of causal effects from observational data. We also introduce the instrumental variable method for inferring causal effects in the presence of hidden variables. In Section 1.5, we introduce the concept of generalizing prediction functions.

1.1. Causal Models

Causal or statistical causal models are enhanced statistical (probabilistic) models which first and foremost specify a probability distribution over a system of random

variables exactly as regular statistical models do. Furthermore, these models are enhanced with a preconceived notion of how the system acts under external manipulation. We further highlight the fundamental differences between statistical and causal models in the next section.

In the rapidly increasing literature on statistical causal modeling, different frameworks exist for defining and manipulating causal models. Some of the more popular frameworks are structural causal models (Pearl, 2009; Peters et al., 2017), causal graphical models (Spirtes et al., 2000), and the potential outcomes framework (Rubin, 1974, 2005). They all render interventional and counterfactual questions well-defined, but since their construction differs, the underlying causal assumptions needed to infer answers to such questions also differs. Thus, depending on the application, one framework may present the causal assumptions in a manner that is more easily digested compared to other frameworks. In this thesis, we work under the framework of structural causal models. We define these models formally in Section 1.1.3.

1.1.1. Statistical and Causal Models

First, consider a statistical model over the random variables X and Y. For example, a typical specification of the association between X and Y in a linear regression model is given by

$$Y = \gamma X + \varepsilon, \tag{1.1}$$

for some $\gamma \in \mathbb{R}$ with X and ε being mutually independent standard normal distributed random variables. This statistical model specifies a simultaneous distribution over (X, Y) given by

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma \\ \gamma & 1+\gamma^2 \end{pmatrix} \right).$$

Given independent and identically distributed (i.i.d.) data generated in accordance with the above specified statistical model, we may consistently estimate the statistical parameter γ by, for example, the ordinary least squares estimator. Knowledge of the statistical model and the statistical parameter γ fully specifies the simultaneous distribution over (X, Y), allowing us to derive predictions for new i.i.d. observations. For example, we may derive the probability that Y is positive given that we have observed that X is positive, or the conditional expectation of Y given an observed value of X, i.e., $E[Y|X = x] = \gamma x$.

The specification of the statistical model in Equation (1.1) may look as if Y is generated by a process that adds noise to γX. In which case, a natural interpretation is that if we were to increase X artificially, we would see an increase in Y if γ is positive. Such interpretations are not valid as a statistical model only specifies an observational distribution. More specifically, the above interpretation relies on a causal assumption of the observed system, i.e., a causal physical

mechanism that outputs Y from the input X and that this physical mechanism does not change when artificially intervening on the input X.

Statistical causal models give us the language and tools to specify and analyze such extended interpretations of statistical models. However, it is worth noting that causal interpretations always require causal assumptions. Without agreeing to certain unfalsifiable causal assumptions, one can never infer causal effects or relations from observational data.

1.1.2. Graphs

Before we define structural causal models, we introduce some graph terminology used throughout this thesis. Graphs are vital in causal reasoning and inference; they allow us to analyze and visualize the causal relations between variables in a system.

A *directed graph* $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (nodes) $V = \{1, \dots, p\}$ and a collection of directed edges $\mathcal{E} \subseteq \{(j \to i) \equiv (j, i) : i, j \in V, i \neq j\}$. We let $\mathrm{pa}^{\mathcal{G}}(i) := \{v \in V : \exists (v, i) \in \mathcal{E}\}$ and $\mathrm{ch}^{\mathcal{G}}(i) := \{v \in V : \exists (j, v) \in \mathcal{E}\}$ denote the *parents* and *children* of node $i \in V$ and we define root nodes $\mathrm{rt}(\mathcal{G}) := \{v \in V : \mathrm{pa}^{\mathcal{G}}(i) = \emptyset\}$ as nodes with no parents (that is, no incoming edges). Two nodes are *adjacent* if there exists an edge between them and a *v-structure* consists three nodes where one node is a child of two non-adjacent nodes. A *path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2, ..., i_k)$ of adjacent nodes, i.e., a sequence of pairs of nodes such that for all $j \in \{1, \dots, k-1\}$, we have either $(i_j \to i_{j+1}) \in \mathcal{E}$ or $(i_{j+1} \to i_j) \in \mathcal{E}$. A *directed path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2, ..., i_k)$ of pairs of nodes such that for all $j \in \{1, \dots, k-1\}$, we have $(i_j \to i_{j+1}) \in \mathcal{E}$. Furthermore, we let $\mathrm{an}^{\mathcal{G}}(i)$ and $\mathrm{de}^{\mathcal{G}}(i)$ denote the *ancestors* and *descendants* of node $i \in V$, consisting of all nodes $j \in V$ for which there exists a directed path to and from i, respectively.

A *directed acyclic graph* (DAG) is a directed graph that does not contain any directed cycles, i.e., directed paths visiting the same node twice. We say that a graph is *connected* if a path exists between any two nodes. A *directed tree* is a connected DAG in which all nodes have at most one parent. More specifically, every node has a unique parent except the root node, which has no parent. The root node $\mathrm{rt}(\mathcal{G})$ is the unique node such that there exists a directed path from $\mathrm{rt}(\mathcal{G})$ to any other node in the directed tree. A directed tree is also called an *arborescence*, a *directed rooted tree* and a *rooted out-tree* in graph theory. We let $\mathcal{T}_p$ denote the set of all directed trees of $p \in \mathbb{N}_{>0}$ nodes. A graph $\mathcal{G}' = (V', \mathcal{E}')$ is a *subgraph* of another graph $\mathcal{G} = (V, \mathcal{E})$ if $V' \subseteq V$, $\mathcal{E}' \subseteq \mathcal{E}$. A subgraph is *spanning* if $V' = V$.

An *undirected graph* $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ nodes (vertices) $V = \{1, \dots, p\}$ and a collection of undirected edges $\mathcal{E} \subseteq \{\{j, i\} : i, j \in V, i \neq j\}$ and a *partially directed graph* or *mixed graph* $\mathcal{G} = (V, \mathcal{E}_u, \mathcal{E}_d)$ has both a collection of undirected edges $\mathcal{E}_u \subseteq \{\{j, i\} : i, j \in V, i \neq j\}$ and a collection of directed edges $\mathcal{E}_d \subseteq \{(j, i) : i, j \in V, i \neq j\}$.

1.1.2.1. D-separation

Pearl's d-separation (Pearl, 2009) is a graphical notion that will allow us to deduce conditional and unconditional independence statements concerning system variables generated by a structural causal model by analyzing the corresponding causal graph. For now, we introduce it as a purely graphical definition concerning directed acyclic graphs. Suppose that we have a directed acyclic graph $\mathcal{G} = (V, \mathcal{E})$. We say that a path $(i_1, ..., i_k)$ in $\mathcal{G}$ between two nodes i_1 and i_k is *blocked* by a collection of nodes $C \subseteq V \setminus \{i_1, i_k\}$ if either

(i) there exists $m \in \{2, ..., k-1\}$ such that $i_m \in C$ and the path contains a subpath of the form $i_{m-1} \to i_m \to i_{m+1}$, $i_{m-1} \leftarrow i_m \leftarrow i_{m-1}$ or $i_{m-1} \leftarrow i_m \to i_{m+1}$, or

(ii) there exists $m \in \{2, ..., k-1\}$ for which neither the node i_m nor any of its descendants are in C, i.e., $(\{i_m\} \cup \mathrm{de}^{\mathcal{G}}(i_m)) \cap C = \emptyset$, and the path contains the subpath $i_{m-1} \to i_m \leftarrow i_{m+1}$.

Definition 1.1 (d-separation)**.** *Consider a directed acyclic graph $\mathcal{G} = (V, \mathcal{E})$. Let $A, B, C \subseteq V$ be three distinct subsets of nodes. A and B are d-separated by C in $\mathcal{G}$, written $A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C$ if and only if all paths between any two nodes in A and B are blocked by C.*

1.1.3. Structural Causal Models

Causal models allow one to specify an observational probability distribution over a system of variables (i.e., a statistical model) but also enable one to reason about interventional and counterfactual questions. This section introduces models with these properties from the framework of structural causal models (SCMs). Later in this section, we introduce interventions, but we refrain from introducing counterfactual reasoning since this thesis does not contribute to this area of research.

Definition 1.2 (Structural causal models)**.** *A structural causal model $M = (Q, \mathcal{S})$ of dimension $p \in \mathbb{N}_{>0}$ consists of a noise distribution Q on $\mathbb{R}^p$ with mutually independent marginals and p structural assignments $\mathcal{S}$:*

$$1 \leq i \leq p: \quad X_i := f_i(X_{\mathrm{PA}(i)}, N_i),$$

where $X_{\mathrm{PA}(i)} \subseteq X = (X_1, ..., X_p)$ denotes the parents or direct causes of X_i and $N = (N_1, ..., N_p) \sim Q$.

The collection of functions $(f_i)_{1 \leq i \leq p}$ and variables $N = (N_1, ..., N_p)$, present in the structural assignments, are called the causal functions and the noise innovations, respectively. Structural causal models are also known as structural equation models or simultaneous equation models in statistics and econometrics (applied with varying degrees of causal interpretation, see, e.g., Pearl, 2012).

We distinguish between two fundamentally different SCM structures; those that are cyclic and those that are acyclic. Whether or not an SCM is cyclic or acyclic plays an essential role in constructing a solution, i.e., the induced random system of variables satisfying the structural assignments.

Definition 1.3 (Acyclic and cyclic SCMs)**.** *A p-dimensional SCM $M = (Q, \mathcal{S})$ is acyclic if there exists a causal order π, i.e., a permutation $\pi : \{1, ..., p\} \to \{1, ..., p\}$, satisfying $\pi(j) < \pi(i)$ whenever $j \in \mathrm{PA}(i)$ for all $1 \leq i \leq p$. An SCM called cyclic if it is not acyclic.*

Let $M = (Q, \mathcal{S})$ be an SCM and let $N : (\Omega, \mathbb{F}) \to \mathbb{R}^p$ and $X : (\Omega, \mathbb{F}) \to \mathbb{R}^p$ be defined on a common probability space $(\Omega, \mathbb{F}, P)$ such that $N \sim Q$. We say that the pair (X, N) solves M if

$$X \overset{\text{a.s.}}{=} f(X, N),$$

where $f(x, n) := (f_1(x_{\mathrm{PA}(1)}, n_1), ..., f_p(x_{\mathrm{PA}(p)}, n_p))$ are the structural assignments $\mathcal{S}$ of M. We say that a random vector X is induced or generated by an SCM $M = (Q, \mathcal{S})$ whenever there exists an $N \sim Q$ such that (X, N) solves the SCM. An SCM-induced random vector is therefore only uniquely defined up to a P-null set.

It is, in general, not guaranteed that solutions exist to cyclic a SCM; see Bongers et al. (2021) for further information on the theoretical foundations of cyclic SCMs. An acyclic SCMs $M = (Q, \mathcal{S})$ is, however, always solvable. Suppose that we have a random vector $N = (N_1, ..., N_p) : (\Omega, \mathbb{F}) \to \mathbb{R}^p$ with $N \sim Q$ and that π is the causal order of the acyclic SCM. We can now define the random vector $X : (\Omega, \mathbb{F}) \to \mathbb{R}^p$ in increasing order of $i \in \{1, ..., p\}$,

$$X_{\pi^{-1}(i)} := f_i(X_{\mathrm{PA}(\pi^{-1}(i))}, N_i),$$

which by definition solves the SCM.

Example 1.1. Consider the acyclic structural causal model given by a noise innovation distribution Q and structural assignments

$$\begin{aligned} X_1 &:= f_1(N_1), \\ X_2 &:= f_2(X_1, N_2), \\ X_3 &:= f_3(X_1, X_2, N_3), \end{aligned}$$

where $N = (N_1, N_2, N_3) \sim Q$. This SCM has a causal order given by

$$(\pi(1), \pi(2), \pi(3)) = (1, 2, 3),$$

so we can given a noise innovation N iteratively define X_1, X_2 and finally X_3. ∘

We define the induced or observational distribution of a solvable SCM M by the push-forward measure $P_X = X(P)$ on $\mathbb{R}^p$ for any solution X. Sometimes we also denote the observational distribution by P_M. The observational distribution is always uniquely defined.

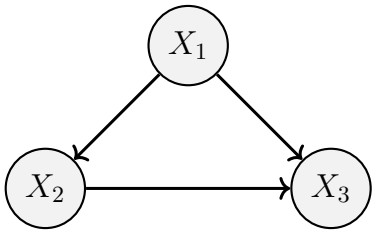


Figure 1.1: The causal graph of the common confounder structural causal model in Example 1.1.

Example 1.2. Consider the SCM of Example 1.1. Now suppose that Q denotes the 3-dimensional standard multivariate normal distribution $N \sim Q = \mathcal{N}(0, I_3)$ and that the structural assignments are linear and given by

$$X_1 := f_1(N_1) \equiv N_1, \quad X_2 := f_2(X_1, N_2) \equiv \alpha X_1 + N_2,$$
$$X_3 := f_3(X_1, X_2, N_3) \equiv \gamma X_1 + \beta X_2 + N_3.$$

By subsitution we find that $X_1 = N_1$, $X_2 = \alpha N_1 + N_2$ and $X_3 = (\gamma + \beta\alpha)N_1 + \beta N_2 + N_3$, from which the induced distribution of M is easily found to be given by $(X_1, X_2, X_3) \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma := \begin{pmatrix} 1 & \alpha & \gamma + \beta\alpha \\ \alpha & \alpha^2 + 1 & \alpha(\gamma + \beta\alpha) + \beta \\ \gamma + \beta\alpha & \alpha(\gamma + \beta\alpha) + \beta & (\gamma + \beta\alpha)^2 + \beta^2 + 1 \end{pmatrix}.$$

∘

Henceforth, we assume that all solvable structural causal models have structurally minimal assignments. That is, for any structural assignment $X_i := f_i(X_{\mathrm{PA}(i)}, N_i)$ there does not exist a $j \in \mathrm{PA}(i)$ and a measurable map $\tilde{f}_i$ such that $f_i(X_{\mathrm{PA}(i)}, N_i) = \tilde{f}_i(X_{\mathrm{PA}(i)\setminus\{j\}}, N_i)$ almost surely.

Definition 1.4 (Causal graph)**.** *The causal directed graph $\mathcal{G} = (V, \mathcal{E})$ of an SCM $M = (Q, \mathcal{S})$ is given by the vertex set $V := \{1, ..., p\}$ and direct edges drawn from each $j \in \mathrm{PA}(i)$ to i for all $i \in V$, i.e.,*

$$\mathcal{E} = \{(j \to i) : i \in V, j \in \mathrm{PA}(i)\}.$$

That is, the causal graph is determined by letting $\mathrm{pa}^{\mathcal{G}}(i) := \mathrm{PA}(i)$ for all $i \in V$.

The causal graph of an acyclic SCM is, therefore, always a DAG. In Figure 1.1, we have illustrated the causal graph of the acyclic structural causal model $M = (Q, \mathcal{S})$ from Example 1.1.

In this thesis, we are mainly concerned with linear cyclic SCMs and general acyclic SCMs. Example 1.3 highlights sufficient conditions for the existence and construction of solutions to linear cyclic SCMs.

Example 1.3 (Linear cyclic SCMs.)**.** A linear cyclic SCM $M = (Q, \mathcal{S})$ satisfies linear structural assignments. That is, for each $1 \leq i \leq p$, the structural assignment is given by

$$X_i := f_i(X_{\mathrm{PA}(i)}, X_i) \equiv b_i^\intercal X_{\mathrm{PA}(i)} + N_i,$$

for some $b_i \in \mathbb{R}^{|\mathrm{PA}(i)|}$. Now let $B \in \mathbb{R}^{p \times p}$ be a constant matrix such that $x = Bx + n$ conforms with the above structural assignments. If $\rho(B)$, the spectral radius of B, is strictly less than one, then we know that $(I - B)$ is invertible. Hence, $x = (I - B)^{-1}n$. Thus, given a noise innovation $N : (\Omega, \mathbb{F}, P) \to \mathbb{R}^p$ with $N \sim Q$, define $X = (I - B)^{-1}N$ and note that (X, N) solves the SCM, since $X = BX + N$ holds P-almost surely. ∘

For any structural causal model, the induced observational distribution satisfies the global Markov property with respect to the causal graph — a one-way connection between the d-separation statements in the causal graph and conditional independencies in the induced distribution.

Theorem 1.1 (Pearl, 2009, Theorem 1.4.1)**.** *Let $X = (X_1, ..., X_p) \in \mathbb{R}^p$ be random vector induced by an acyclic structural causal model M with acyclic causal graph $\mathcal{G} = (V, \mathcal{E})$. The induced distribution P_X satisfies the global Markov property with respect to the causal graph. That is,*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C \implies X_A \perp\!\!\!\perp X_B \,|\, X_C,$$

for all disjoint subsets $A, B, C \subseteq V = \{1, ..., p\}$.

Thus, the causal graph yields through d-separation a visual representation of conditional independence statements in the observational distribution of a structural causal model.

1.1.3.1. Interventions

So far, the structural causal models only induce an observational distribution, i.e., a statistical model which only allows us to ask and answer questions about probabilistic associations. The main difference between a statistical model and a causal model is the ability to explain the behavior of a stochastic system of variables under external manipulation (intervention). In the search for a tractable behavior of systems under manipulation, one usually assumes autonomy, also called modularity, of the causal (physical) mechanisms of the system we are modeling.

Assumption 1.1 (Autonomy of causal mechanisms; Peters et al., 2017)**.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

The assumption of autonomous causal mechanisms yields the ability to conduct external manipulations of the generative process in selected parts of a system without affecting the generative processes of the remaining system.

Example 1.4 (Autonomy in a cause-effect system)**.** Consider a bivariate cause-effect system where X causes Y. Suppose that f is the mechanism that produces Y given the cause/input X, i.e., $Y := f(X)$. Assumption 1.1 translates to independence between cause and mechanism. The assumption of autonomous causal mechanisms stipulates that any external manipulation of X does not affect the mechanism f, which produces Y. ∘

The assumption of autonomous causal mechanisms allows us to analyze the behavior of a system under external interventions in a tractable fashion.

Definition 1.5. *An intervention i is a map between structural causal models*

$$M = (Q, \mathcal{S}) \overset{i}{\mapsto} (Q^i, \mathcal{S}^i),$$

where Q^i and $\mathcal{S}^i$ are the post-intervention noise distribution and structural assignments. We let $M(i) = (Q^i, \mathcal{S}^i)$ denote the post-intervention structural causal model.

In this introduction, we only concern ourselves with fairly simple interventions. Later chapters will introduce more general interventions as needed. For example, an intervention on a single system variable amounts, by Assumption 1.1, to only changing the structural assignment of said variable; see Example 1.5 below.

Example 1.5. Consider the SCM $M = (Q, \mathcal{S})$ of Example 1.1. Let i be an intervention that randomizes X_2, i.e., forces it to obey a distribution P^i independently of the outcome of its original direct cause X_1. That is, we change the structural assignments in the following way:

$$\mathcal{S} = \begin{cases} X_1 := f_1(N_1), \\ X_2 := f_2(X_1, N_2), \\ X_3 := f_3(X_1, X_2, N_3), \end{cases} \quad \overset{i}{\longmapsto} \quad \mathcal{S}^i = \begin{cases} X_1 := f_1(N_1), \\ X_2 := \tilde{N}_2, \\ X_3 := f_3(X_1, X_2, N_3), \end{cases}$$

where $(N_1, N_2, N_3) \sim Q = Q_1 \times Q_2 \times Q_3$ and $(N_1, \tilde{N}_2, N_3) \sim Q^i = Q_1 \times P^i \times Q_3$. In Figure 1.2, we have illustrated the corresponding changes to the causal graph. The edge from X_1 to X_2 is removed due to the effect breaking intervention. ∘

Interventions need not break the direct link of the original causes; it can also simply change the causal mechanism which produces the variable from its causes. We denote such interventions on single system variables, say, X_i, by

$$\mathrm{do}(X_i := \tilde{f}_i(X_{\widetilde{\mathrm{PA}}(i)}, \tilde{N}_i)),$$

where $\tilde{f}$ is a (possibly) new causal mechanism taking the new direct causes $\widetilde{\mathrm{PA}}(i)$ and noise innovation $\tilde{N}_i$ as inputs. For example, the intervention in Example 1.5 is denoted by $\mathrm{do}(X_2 := \tilde{N}_2)$ with $\tilde{N}_2 \sim P^i$. In the upcoming chapters, we use slightly different notations for intervention-induced distributions, i.e., the post-intervention simultaneous distribution of the system. For example, the intervention-induced

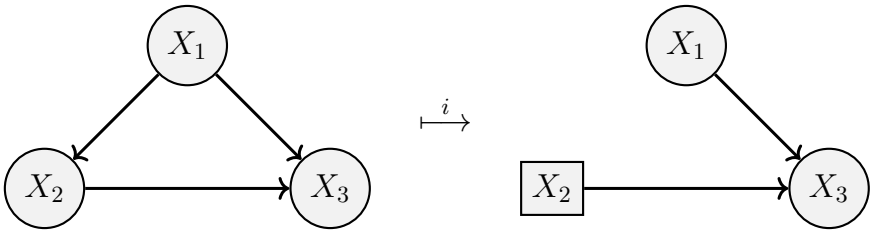


Figure 1.2: Illustration of the original and post-intervention causal graph for the structural causal model and intervention considered in Example 1.5.

distribution for the intervention $i = \mathrm{do}(X_2 := \tilde{N}_2)$ in an SCM M may be denoted by

$$P_{M(i)}, \quad \text{or} \quad P_M^{\mathrm{do}(X_2:=\tilde{N}_2)}, \quad \text{or} \quad P^{\mathrm{do}(X_2:=\tilde{N}_2)},$$

depending on whether or not the underlying SCM M and intervention i is clear from the context. In the example below, we derive an intervention-induced distribution.

Example 1.6. Consider the SCM $M = (Q, \mathcal{S})$ of Example 1.2. Suppose that we conduct the intervention $i = \mathrm{do}(X_2 := \tilde{N}_2)$ with $\tilde{N}_2 \sim \mathcal{N}(0, 1)$ independent from the original noise innovations of the system. The post-intervention structural assignments are now given by

$$X_1 := N_1, \quad X_2 := \tilde{N}_2, \quad X_3 := \gamma X_1 + \beta X_2 + N_3.$$

Thus, $X_1 = N_1$, $X_2 = \tilde{N}_2$ and $X_3 = \gamma N_1 + \beta \tilde{N}_2 + N_3$, so the intervention-induced distribution is given by $P_{M(i)} = \mathcal{N}(0, \Sigma)$ where

$$\Sigma := \begin{pmatrix} 1 & 0 & \gamma \\ 0 & 1 & \beta \\ \gamma & \beta & \gamma^2 + \beta^2 + 1 \end{pmatrix}.$$

∘

1.2. The Difficulties of Causal Inference

Inferential targets in causal models can be statistical or causal quantities. For example, we may be interested in statistical targets, i.e., quantities defined in terms of the joint distribution of the system variables. Statistical targets include, for example, the correlation between variables, conditional probabilities, or conditional expectations between certain variables. Causal targets are non-statistical quantities defined in terms of a causal model (Pearl, 2009). Common causal targets include the causal graph (or parts thereof, e.g., the direct causes of a specific variable),

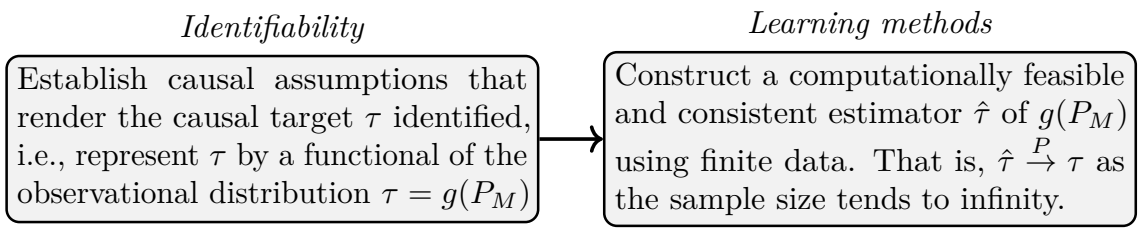


Figure 1.3: Flowchart of causal inference from observational data.

causal effects, and general post-interventional probabilistic quantities of system variables, i.e., the post-intervention distribution or a derivative thereof.

However, as causal quantities are not defined in terms of the system's observational distribution, their inference from observational data will instead rely on causal assumptions about the system of interest. Such assumptions are, by definition, not falsifiable by observational data and therefore purely rests on the practitioner's expert judgment (Pearl, 2009).

There are two main aspects to learning causal targets: identifiability and learning methods; see the flowchart in Figure 1.3. First, we have the aspect of identifiability; see Section 1.2.1. Here we are concerned with the theoretical ability to infer the target from the observational distribution of the system. Second, in the affirmation of identifiability, we have the aspect of constructing learning methods (identification); see Section 1.2.2. Here we are concerned with estimating the causal target from finite data, similar to regular inference of statistical quantities.

1.2.1. Identifiability

In practice, most causal targets can be recovered by conducting specific interventions in a system and analyzing the observed changes. For example, it is possible to recover the average treatment effect of a drug by conducting a randomized controlled trial (Peirce, 1883) where one randomly assigns a patient the treatment or a placebo. The random assignment can be seen as an intervention in which the treatment indicator (i.e., whether the patients get the drug or a placebo) is externally manipulated to follow the outcome of a binary random variable that is independent of other system variables (e.g., patient covariates, etc.). However, due to either ethical, monetary or practical reasons, we may not be able to conduct the preferred system interventions that would enable us to quantify the causal targets. In this thesis, we are mainly concerned with the latter scenario where interventions are not possible.

In theory, there could be several distinct data-generating processes (causal models) that are observationally equivalent (induces identical observational distributions) but differ on the causal quantity of interest. Hence, an essential aspect of causal modeling is specifying causal assumptions that allow us to infer the causal targets from the observational distribution alone. A causal target is said to be

identified if we can theoretically infer it from the observational distribution.

A lot of causal targets become identified once the causal graph of the causal model is known. Thus, we either have to resort to expert judgment on the causal structure or infer the structure from data. In Section 1.3, we highlight some standard structure learning methods and detail the causal assumptions they rely on.

1.2.2. Learning Methods

The next problem in causal inference is inferring or learning the causal target of interest from finite data in a consistent and computationally feasible way. In the affirmation of identifiability, we know that the observational distribution uniquely determines the causal target. Thus, in theory, we could infer the causal target given complete knowledge of the observational distribution.

Under appropriate causal assumptions, some causal targets are given by quantities of the observational distribution (distributional features) for which inference has been well-studied in the statistical literature, e.g., conditional expectations or linear regression coefficients. In such cases, inference can be achieved by simply applying established statistical inference methods. However, sometimes the causal target is not a commonly studied quantity of the observational distribution. In these cases, inference requires new methods with accompanying theoretical large sample guarantees.

1.3. Learning Causal Graphs

The causal graph of a causal model is often of interest to practitioners due to the intrinsic value of knowing what system components cause a specific variable. Alternatively, one is interested in the causal structure since other causal targets become identified from the observational distribution once the causal graph is known; see, e.g., Section 1.4.

We focus on the problem of inferring the causal structure from observational data. However, as we have previously mentioned, inference of causal quantities from observational data necessitates causal assumptions on the system of interest. That is, we need causal assumptions that make it theoretically possible to infer the causal graph of an acyclic SCM from its induced distribution.

Standard structure learning methods are classified as independence-based (also known as constraint-based), score-based, or mixed. Structure learning methods that are independence-based rest on the nonparametric causal assumption of faithfulness; see Definition 1.6. Faithfulness renders parts of the causal structure identified through the independence constraints encoded in the observational distribution. On the other hand, score-based methods rest on causal assumptions on the causal mechanisms and noise innovations of the system of interest.

In Section 1.3.1, we introduce the causal assumptions for independence-based

structure learning and briefly discuss established methods for inference. Section 1.3.2 introduces score-based approaches to causal structure learning, which is also the topic of Chapter 4.

1.3.1. Independence-based Structure Learning

Independence-based structure learning methods infer parts of the causal graph by utilizing (conditional) independence constraints encoded in observational distribution. We have previously seen that the induced distribution of an acyclic SCM is Markov with respect to the causal graph. However, for learning the structure itself, this is a useless property as, for example, any SCM induced distribution is also Markov with respect to the fully connected graph. In general, without further causal assumptions, the (conditional) independence constraints encoded in the observational distribution do not yield any causal graph information. This problem leads us to the fundamental causal assumption on which independence-based structure learning methods rests; the assumption of faithfulness with respect to the causal graph.

Definition 1.6 (Faithfulness)**.** *Let $X = (X_1, ..., X_p) \in \mathbb{R}^p$ be a random vector with distribution P_X and let $\mathcal{G}$ be a DAG with nodes $V = \{1, ..., p\}$. The distribution P_X is said to be faithful with respect to the graph $\mathcal{G}$ if*

$$X_A \perp\!\!\!\perp X_B | X_C \implies A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C$$

for all disjoint subsets $A, B, C \subseteq V = \{1, ..., p\}$.

Thus, if we assume that the induced distribution of an acyclic SCM is faithful to the causal graph, then by the global Markov property, we have a one-to-one correspondence between d-separations in the causal graph and conditional independence constraints encoded by the induced distribution. Independence-based structure learning methods exploit this correspondence: utilizing conditional independence testing, one draws inference on conditional independence statements that allow one to draw inference about the causal graph through the faithfulness assumption.

Faithfulness implies causal minimality (Peters et al., 2017, Proposition 6.35), i.e., if P_X is faithful with respect to the causal graph $\mathcal{G}$, then P_X it is not Markov with respect to any proper subgraph of $\mathcal{G}$. Faithfulness is a causal assumption that is not satisfied in general; see Example 1.7 below.

Example 1.7. Consider the linear Gaussian SCM of Example 1.2, with causal graph is illustrated in Figure 1.1. The structural assignments are given by

$$X_1 := N_1, \quad X_2 := \alpha X_1 + N_2, \quad X_3 := \gamma X_1 + \beta X_2 + N_3,$$

where $N = (N_1, N_2, N_3) \sim \mathcal{N}(0, I)$. If $\alpha\beta = -\gamma$, then $X_2 \perp\!\!\!\perp X_3$. However, X_2 is not d-separated from X_3 given the empty set, so faithfulness is not satisfied with respect to the causal graph. ∘

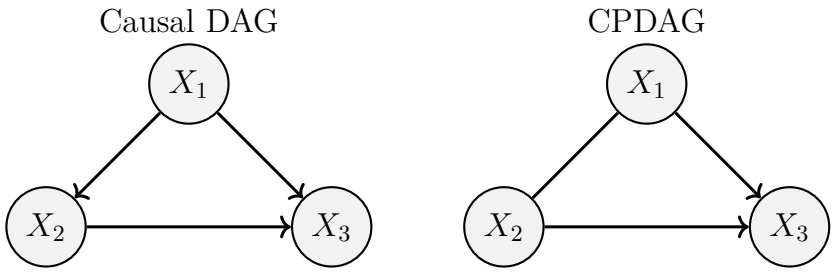


Figure 1.4: The causal graph from the SCM of Example 1.1 and the corresponding CPDAG representing its Markov equivalence class.

Let us discuss what parts of the causal structure the assumption of faithfulness identifies. That is, we will discuss what it entails that we can infer all d-separation statements of a causal graph. To this end, we say that two graphs $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are Markov equivalent if every probability distribution that is globally Markov with respect to $\mathcal{G}$ is also globally Markov with respect to $\tilde{\mathcal{G}}$ and vice versa. The Markov equivalence class (MEC) of a graph $\mathcal{G}$, $\text{MEC}(\mathcal{G})$, consists of all graphs that are Markov equivalent to $\mathcal{G}$. It has been shown that $\text{MEC}(\mathcal{G}) = \{\tilde{\mathcal{G}}$ is a DAG : $\tilde{\mathcal{G}}$ and $\mathcal{G}$ share the same d-separations$\}$ (Verma and Pearl, 1990b), so faithfulness implies that the Markov equivalence class of the causal graph is identified. Finally, the following theorem quantifies the shared structure of all DAGs in the Markov equivalence class.

Theorem 1.2 (Verma and Pearl, 1990a)**.** *Two DAGs are Markov equivalent if and only if they share the same skeleton and v-structures.*

Thus, it is possible to represent the Markov equivalence class of a DAG $\mathcal{G}$ by a unique partially directed acyclic graph (PDAG) known as the *completed* PDAG (CPDAG) with the skeleton and directed edges that make up v-structures shared by all members. In Figure 1.4, we have illustrated the causal graph and the corresponding CPDAG representing its Markov equivalence class of the SCM from Example 1.1.

As for learning the CPDAG, we can use the popular PC-algorithm (Spirtes et al., 2000). The contributions in this thesis do not add to the literature on independence-based structure learning, so we refer to Spirtes et al. (2000) for further details on the algorithm. Nevertheless, given oracle knowledge on conditional independence statements, the PC-algorithm recovers the CPDAG whenever faithfulness is satisfied. However, when inferring the CPDAG from finite data, the conditional independence statements have to be inferred by successive conditional independence tests. One usually chooses a fixed significance level for the tests, but due to the successive testing, one loses the error quantification of the method as a whole. Furthermore, conditional independence tests can not have power against any alternative (Shah and Peters, 2020) unless specific distributional assumptions are made, such as joint Gaussianity. Type I errors of the conditional independence

tests can lead to the removal of causal edges and the inclusion of non-causal edges in the resulting CPDAG (Spirtes et al., 2000).

1.3.2. Score-based Structure Learning

Score-based approaches to causal structure learning use (parametric) assumptions on the structural causal model that allow for the construction of a scoring function for causal structures. That is, in the affirmation of identifiability of the causal graph (or parts thereof), we define a (population) score function ℓ that only attains its minimum in the causal graph

$$\mathcal{G} = \underset{\tilde{\mathcal{G}}:\tilde{\mathcal{G}} \text{ is a DAG}}{\arg\min} \ \ell(\tilde{\mathcal{G}}). \tag{1.1}$$

The greedy equivalence search (GES, Chickering, 2002) assumes faithfulness which renders the MEC identified. Under the additional assumption of joint Gaussianity of the observed distribution, GES minimizes a BIC-penalized likelihood score function directly on the space of Markov equivalence classes.

Causal system assumptions that guarantee identifiability of the causal graph itself have also been studied. For example, in SCMs with additive Gaussian noise and nonlinear causal functions, the causal graph is identified; see the introduction of Chapter 4 for an overview. However, in the pursuit of the causal graph, we stumble onto new computationally problematic issues. Even though the optimization problem in Equation (1.1) is guaranteed to have a unique minimum, the optimization problem is a combinatorial problem with a search space cardinality that grows super-exponentially in the number of system variables. Thus, for even moderately large systems, brute-force optimization (exhaustive search) becomes computationally infeasible.

At the current state of the literature, no optimization procedure guarantees to solve the problem with computationally feasible time complexity for large systems. However, several heuristic optimization procedures have been proposed. For example, Bühlmann et al. (2014) propose a greedy search technique on the space of DAGs, and Zheng et al. (2018) propose an equivalent continuous albeit non-convex representation of the optimization problem in Equation (1.1). These approaches do not guarantee to recover the causal graph. For example, the non-convex continuous optimization problem representation necessitates naive optimization approaches with no guarantees of not getting stuck in a local minimum. Moreover, it is currently being discussed whether the seemingly remarkable performance in simulation studies of Zheng et al. (2018) is due to the exploitation of simulated DAG artifacts rather than successful naive optimization; see Reisach et al. (2021) and Section 1.3.3.1 below. In Section 1.3.2.2, we show an example where the greedy search of Bühlmann et al. (2014) fails. Hence, there is currently no practical method that guarantees the recovery of the actual causal graph with probability tending to one in the large sample limit.

1.3.2.1. Causal Structure Learning for Directed Trees

In Chapter 4, we take a slightly different approach to the computational problems associated with recovering the actual causal graph in score-based approaches. Instead of proposing another heuristic optimization procedure, we look at what relaxations in the system complexity allow for exact score-function minimization.

In particular, we restrict our attention to less complex systems with causal graphs given as directed trees and additive noise. While brute-force minimization over the space of directed trees is still computationally infeasible, i.e., the search space still grows super-exponentially in the system size, we show that the optimization is possible with polynomial time complexity. More specifically, we show that Chu–Liu–Edmonds' algorithm (proposed independently by Chu and Liu, 1965; Edmonds, 1967) from graph theory solves the optimization problem.

We show that the proposed method, called causal additive trees (CAT), is consistent under weak conditions. Moreover, due to the reasonably simple causal structure, we provide inference results to test causal substructure hypotheses. Our proposed hypothesis testing procedure retains its level-guarantees under post-selection hypothesis generation and multiple testing. Furthermore, we investigate the identifiability gap, i.e., the minimum score difference between the causal graph and any alternative graph. For Gaussian noise innovations, we provide a lower bound that depends only on local dependence properties. That is, the identifiability of the causal graph reduces to a purely local property for Gaussian additive noise models.

1.3.2.2. When Greedy Searches Fail

Greedy search techniques do not, in general, come with theoretical guarantees. We now present an example where the greedy search of Bühlmann et al. (2014), called CAM, consistently fails to recover the causal graph, while our method CAT successfully recovers the causal graph as the sample size increases. The following model is taken from Peters et al. (2022). Consider the following three node Gaussian additive structural causal model with causal graph ($X \to Y \to Z$):

$$X := N_X, \quad Y := \frac{X^3}{\mathrm{Var}(X^3)} + N_Y, \quad Z := Y + N_Z, \tag{1.2}$$

where $N_X \sim \mathcal{N}(0, 1.5)$, $N_Y \sim \mathcal{N}(0, 0.5)$ and $N_Z \sim \mathcal{N}(0, 0.5)$ are mutually independent. Our method CAT has two variants: CAT.G and CAT.E using a Gaussian and entropy scoring function, respectively. We simulate data from this model and estimate the causal graph by CAT.G, CAT.E and CAM. Figure 1.5 illustrates the results. Even with increasing sample size, CAM does not converge to the correct answer. The reason is that it selects the wrong edge in the first step of the greedy search algorithm.

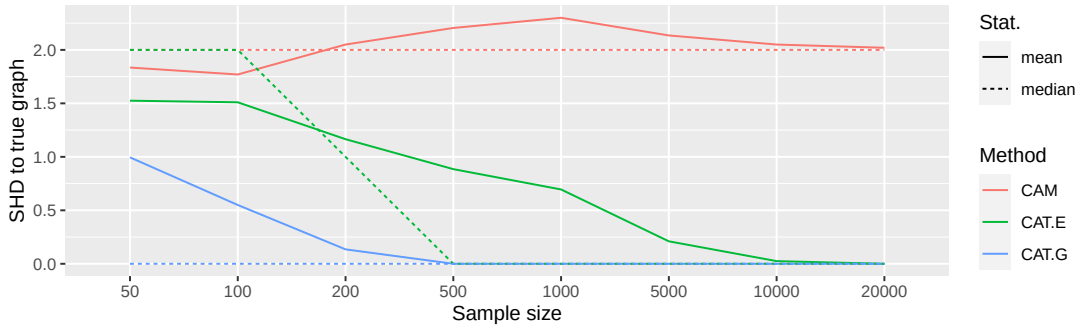


Figure 1.5: Structural hamming distance (SHD, Tsamardinos et al., 2006) performance of CAT.G, CAT.E and CAM in the three node setup of Equation (1.2). The solid and dashed lines represent the mean and median SHD, respectively, based on 200 repetitions.

We now highlight why the greedy search fail. The following explanation relies on the theory presented in Chapter 4, but for now it suffices to know that the score function evaluated in a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{E}}, V)$ is given by the sum of certain edge weights $w_{\mathrm{G}}(j \to i)$ for all edges in the graph. The greedy search technique of CAM iteratively selects the lowest scoring directed edge under the constraint that no cycles is introduced in the resulting graph. Figure 1.6 shows the estimated Gaussian edge weights. The smallest edge weight is given by the wrong edge $(Z \to Y)$ so the greedy search erroneously picks this edge. However, Chu–Liu–Edmonds' algorithm used by CAT correctly realizes that the full score of the correct graph $X \to Y \to Z$ is smaller than the full score of $Z \to Y \to X$ which is recovered by CAM.

1.3.3. Learning Summary Graphs of Time Series

In Chapter 5, we consider the problem of learning summary graphs of time-homogeneous stochastic processes. The paper is the culmination of the authors' participation and victory in the NeurIPS 2019 Causality 4 Climate (C4C) competition.[1] Here, teams were given finite sample data of different simulated d-dimensional time series and then tasked with inferring the underlying summary graph. The summary graph is a simplification of the (infinite) causal graph. It consists of d nodes with an edge from node j to node i if and only if any past values of the j'th coordinate process enter the structural assignment of the i'th coordinate process. For each data set, the participants could upload a weighted adjacency matrix A corresponding to the summary graph where each entry held the belief or score that an edge is present. The online platform, to which the weighted adjacency matrix was uploaded, then scored the method by the area under the curve of the receiver operating characteristic (AUC-ROC) metric.

[1] https://causeme.uv.es/neurips2019

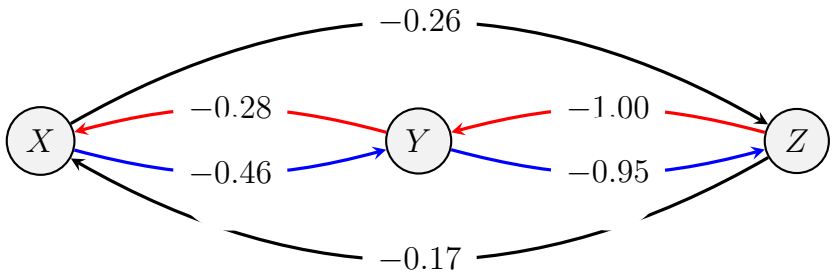


Figure 1.6: Visualization of the edge weights of the experiment in Section 1.3.2.2. Each edge label is the estimated Gaussian edge weight as produced by the CAM scoring method based on 1000000 i.i.d. observations generated from the structural causal system of Equation (1.2). The red edges are recovered by the greedy search of CAM and the blue edges are recovered by Chu–Liu–Edmonds' algorithm of CAT. We see that $-1.41 = \hat{w}_{\mathrm{G}}(X \to Y) + \hat{w}_{\mathrm{G}}(Y \to Z) < \hat{w}_{\mathrm{G}}(Z \to Y) + \hat{w}_{\mathrm{G}}(Y \to X) = -1.28$.

The receiver operating characteristic is a function $\mathrm{ROC} : [0,1] \to [0,1]^2$ which for a binary classifier system takes a threshold $t \in [0,1]$ and yields $\mathrm{ROC}(t) = (\mathrm{FPR}(t), \mathrm{TPR}(t))$ where $\mathrm{FPR}(t)$ and $\mathrm{TPR}(t)$, are the false positive rate and true positive rate of the classifier system using a threshold of t. In our setting, for a fixed threshold $t \in [0,1]$, we convert the weighted adjacency matrix A to a binary adjacency matrix $A^*(t)$, where $A^*(t)_{ji} = 1_{[A_{ji}/\max_{ji} A_{ji}, 1]}(t)$. The true positive rate (TPR) using the threshold t is then given by calculating the fraction of correct edges in $A^*(t)$ over the number of true edges in the underlying summary graph. The false positive rate (FPR) is given by the number of incorrect edges in $A^*(t)$ over the total number of absent edges in the underlying summary graph.

In the paper, we detail our algorithms and present heuristic justifications for our choices. Two important observations are that: 1) our methods using linear regression to capture causal effects seems to work well even though the true causal mechanisms are nonlinear, and 2) the size of the estimated linear coefficients seemed to work better than using an associated test-statistics for a test of vanishing linear effect. We now present a heuristic justification for why linear methods can still be used to discover nonlinear causal effects. In Section 1.3.3.1, we discuss why using the size of linear regression coefficients can outperform methods using corresponding test sizes for tests of vanishing linear effect.

Consider a simple (single-lag) time-homogeneous discrete-time stochastic process $(X(t))_{t \in \mathbb{N}_+}$, where for each time step $t \geq 1$ the process $X(t) \in \mathbb{R}^d$ is driven by past values according to $X(t) := F(X(t-1)) + N(t)$, for $t \geq 1$, some fixed function $F = (F_1, ..., F_d) : \mathbb{R}^d \to \mathbb{R}^d$, noise innovations $(N(t))_{t \geq 1}$ and some initial distribution X_0. As such consider the parameter $\theta_{ji}(t) = \mathbb{E}|\partial_j F_i(X(t))|$. When

the process $(X(t))_{t\in\mathbb{N}_+}$ is strictly stationary, this parameter does not depend on t, and it is clear that when there is an edge in the summary graph from j to i, then $\theta_{ji} > 0$ and $\theta_{ji} = 0$ otherwise. In order to detect regions with non-zero gradients of F, we create random bootstrap samples $\mathcal{D}_1, ..., \mathcal{D}_B$ of the observed time series. We then obtain (possibly penalized) linear regression coefficients each bootstrap sample. The idea is that, if there is no link in the summary graph, then all the bootstrap coefficients are likely small. On the other hand, if $\theta_{ji} > 0$, then there might be at least one large absolute coefficient. We then use the average of the absolute regression coefficients over the B bootstrap samples as a proxy for θ_{ji}. We average the absolute coefficients to avoid possible cancellation. This estimate does not contain any information about whether there is a positive or negative effect from $X_j(t-1)$ to $X_i(t)$, nor can it be used for prediction purposes. It solely serves as a score or belief in the existence of a cause-effect mechanism between past values of X_j onto X_i.

1.3.3.1. Artifacts in DAG Models

In the above learning framework we were only interested in the belief of a causal link, i.e., only quantifying that a linear coefficient is nonvanishing. An immediate question is now: why do we not use, for example, the T-statistic corresponding to the test for the hypothesis that the regression coefficients are zero instead of the absolute size of the corresponding coefficient? The answer is that our proposed algorithms are to some extend tailored towards maximizing the AUC-ROC on the simulated time series data. We explicitly saw a drop in performance when changing to test statistics or p-values. As shown in the simulation experiment, such behavior is also seen in general DAG models where the marginal variance tends to increase the further down the causal order we go.

We exploited this in our methods, but this is not a desirable feature of general-purpose structure learning algorithms, since we generally have no evidence or a priori belief that real-world systems exhibit such behavior. Reisach et al. (2021) further investigated these observations. They argue that for simulated linear additive noise DAG models, it is very easy to, unknowingly, construct models for which the marginal variance increases with the causal order. For example, they show that the benchmark setup of, e.g., Zheng et al. (2018) and Ng et al. (2020) is highly affected by this increasing variance artifact. The problem with such benchmarking setups is that heuristic score-based approaches like Zheng et al. (2018) can exhibit remarkable performance that is superior to other more canonical and well-studied structure learning methods. This performance superiority is immediately lost when data is properly standardized.

1.4. Learning Causal Effects

The previous section discussed causal structure learning methods that enable us to learn the existence of cause-effect relationships in stochastic systems. We

may also be interested in knowing how a system variable behaves under external manipulation (interventions) on the causes of said variable.

Consider, for example, a binary treatment indicator $T \in \{0, 1\}$ indicating whether a patient is administered a specific treatment or not. Suppose that we want to quantify the effect of said treatment on a response variable Y, e.g., a post-treatment indicator of a specific disease or some other biochemical marker of interest. One way to quantify this effect is to consider the treatment's average causal effect (or average treatment effect) on the response variable. That is, we may consider the difference in the expected response variable under two different interventions:

$$\mathrm{ATE} := \mathbb{E}^{\mathrm{do}(T:=1)}[Y] - \mathbb{E}^{\mathrm{do}(T:=0)}[Y].$$

We may also be interested in quantifying how much a response variable Y is affected by interventions on a continuous system variable X. For example, the expected behavior of Y under interventions that fix X at specific values, i.e., the function $x \mapsto \mathbb{E}^{\mathrm{do}(X:=x)}[Y]$ or its derivative $x \mapsto D_x \, \mathbb{E}^{\mathrm{do}(X:=x)}[Y]$. These quantities provide information about whether the response, on average, will decrease or increase due to applying external manipulation, which artificially increases the continuous variable X.

For certain models where X is a direct cause of the response Y the inferential target quantifying the causal effects becomes the causal coefficients (in linear SCMs) and causal functions (in nonlinear SCMs) appearing in the structural assignments of Y; see Example 1.8 and Example 1.9 below.

Example 1.8 (Causal effects in linear models)**.** Consider a linear additive structural causal model $(Q, \mathcal{S})$ over (Y, X, H) with $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$ and $H \in \mathbb{R}^r$ with structural assignments given by

$$\begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} := \begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} B + N^\intercal,$$

for some strictly lower triangular constant matrix B and noise innovation vector $N \sim Q$ with zero mean. Assume w.l.o.g. that the first column of B is given by $(0, \gamma, \delta)$ such that the structural equation of Y becomes $Y := \gamma^\intercal X + \delta^\intercal H + N_Y$. Since B is strictly lower triangular, we know that the variables H act as possible confounders of the causal effect from X to Y, i.e., the causal effect is not mediated by H. As such, they are unaffected by interventions on X. Now consider the intervention $\mathrm{do}(X := x)$ for some constant $x \in \mathbb{R}^d$ and note that $\mathbb{E}^{\mathrm{do}(X:=x)}[Y] = \mathbb{E}^{\mathrm{do}(X:=x)}[\gamma^\intercal x + \delta^\intercal H + N_Y] = \gamma^\intercal x$. Hence, the causal effect $D_x \, \mathbb{E}^{\mathrm{do}(X:=x)}[Y] = \gamma$, is constant and given by the structural parameters γ. ∘

Example 1.9 (Causal effects in nonlinear additive models)**.** Consider a possibly nonlinear structural causal model $(Q, \mathcal{S})$ over (Y, X, H) with $Y \in \mathbb{R}, X \in \mathbb{R}^d$ and $H \in \mathbb{R}^r$ with the structural assignments given by

$$Y := f(X) + g_1(H, N_Y), \quad X := g_2(H, N_X), \quad H := N_H,$$

for some functions $f : \mathbb{R}^d \to \mathbb{R}$, $g_1 : \mathbb{R}^r \to \mathbb{R}$, $g_2 : \mathbb{R}^r \to \mathbb{R}^d$. Now notice that $\mathbb{E}^{\mathrm{do}(X:=x)}[Y] = f(x) + \mathbb{E}[g_1(H, N_Y)]$ from which we get that $D_x\mathbb{E}^{\mathrm{do}(X:=x)}[Y] = D_x f(x)$. Thus, the problem reduces to finding $x \mapsto D_x f(x)$ or $x \mapsto f(x)$, i.e., the causal function f.

○

Causal effects (and other causal targets) are given by distributional features of post-interventional distributions. Hence, inference should be possibly by observing said interventions and analyzing the resulting data. However, given sufficient knowledge of the causal structure it is possible, in certain settings, to infer the interventional distribution (and derivatives thereof) from the observational distribution. For example, if we in Example 1.9 have that $X := g_2(N_X)$, i.e., that X and Y are not confounded, then intervening coincides with conditioning. That is, the inferential target reduces to $\mathbb{E}^{\mathrm{do}(X:=x)}[Y] = \mathbb{E}[Y|X = x]$, for which inference from observational data is a well-studied statistical problem. The next section introduces adjustment formulas that allow for a similar translation when X and Y are confounded.

1.4.1. Adjustment Formulas

Adjustment formulas allow one to derive intervention distributions in terms of the observational distribution, given that we have sufficient knowledge of the underlying causal structure of the system. The adjustment formulas are known in the different causal modeling frameworks as truncated factorization (Pearl, 2009), the G-computation formula (Robins, 1986), and the manipulation theorem (Spirtes et al., 2000). If all relevant densities exist, we say that a set of variables Z is a valid adjustment set for the causal effect from X to Y if it holds that

$$p_Y^{\mathrm{do}(X:=x)}(y) = \int p_{Y|X,Z}(y|x, z)p_Z(z)\, dz,$$

where $p_Y^{\mathrm{do}(X:=x)}$ is the post-intervention density of Y under the intervention $\mathrm{do}(X := x)$, p_Z is a density of Z and $p_{Y|Z,X}$ is a conditional density of Y given Z and X, both under the observational distribution. Thus, a valid adjustment set allows for the interventional distribution to be represented solely by the observational distribution. Various graphical criteria exist to check whether a set Z is a valid adjustment set for the causal effect from X to Y. For example,

- *Parent adjustment:* Suppose that Y is not a parent of X, $Y \notin \mathrm{PA}(X)$. It holds that the collection of all parents of X, $Z := \mathrm{PA}(X)$, is a valid adjustment set.
- *Backdoor adjustment:* Suppose that Z does not contain X or Y and that (i) Z contains no descendant of X and (ii) Z blocks (see, Section 1.1.2.1) all paths between X and Y with an edge incoming edge into X.

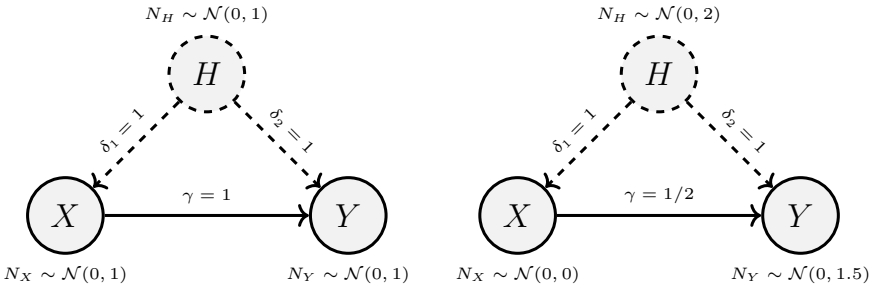


Figure 1.7: Specifications of two linear structural causal models with the same causal graph but different causal coefficients and noise innovation variances; see Example 1.10. In the two linear SCMs, the causal effect from X to Y differs, but the induced observational distributions over (X, Y) coincide.

See Peters et al. (2017) for further characterizations of valid adjustment sets. However, whenever there are hidden (latent) variables, i.e., variables present in the system but not observed, we might not be able to find a valid adjustment set. We discuss this further in the next section.

1.4.2. Inference in the Presence of Hidden Variables

Latent variables further complicates the inference of causal effects. That is, the valid adjustment sets may overlap with the latent variables rendering the use of adjustment formulas to compute causal effects infeasible. In fact, the presence of hidden variables might render the causal effect unidentified. Even when the causal structure and the form of the structural assignments are known a priori, there might be multiple distinct structural causal models that generate identical observational distributions over the observed variables; see Example 1.10.

Example 1.10 (Hidden confounding models)**.** Consider a linear SCM M over (Y, X, H) with $Y \in \mathbb{R}, X \in \mathbb{R}$ and $H \in \mathbb{R}$ where the H denotes a hidden variable, i.e., a variable which can not be observed. Suppose that the structural assignments are given by

$$Y := \gamma X + \delta_1 H + N_Y, \quad X := \delta_2 H + N_X, \quad H := N_H$$

N_Y, N_X, N_H being mutually independent noise innovations. In Figure 1.7, two structural causal models with the above structural assignments are specified. They induce the same observational distribution over X and Y, but the causal effects from X to Y differ. This example clearly illustrates that the causal effect γ is not identified, as it is impossible to infer it from the observational distribution. ∘

In the presence of hidden confounding, we may still be able to identify causal effects by the instrumental variable method.

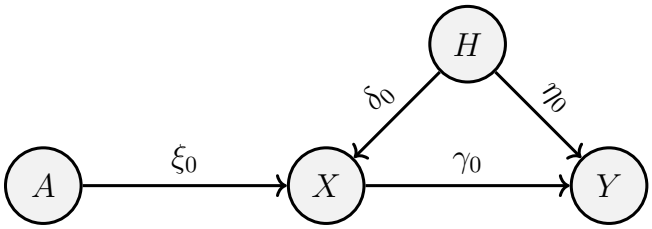


Figure 1.8: The causal graph of the one-dimensional instrumental variable setup.

1.4.2.1. The Instrumental Variable Method

The instrumental variable method (Theil, 1953; Wright, 1928) is a method for identifying and estimating causal effects in the presence of hidden confounding. Suppose that we want to estimate the causal effect from X to Y. The method assumes the existence of system variables A, called instruments, which satisfies the following two criteria (Pearl, 2009):

(i) *Relevance:* A is dependent on the predictors X.

(i) *Exogeneity*: A is independent of all variables (including noise innovations) that influence Y which is not mediated by X. That is, A is independent of Y when X is held fixed: $A \perp\!\!\!\perp Y$ under distributions induced by interventions of the form $\mathrm{do}(X := x)$ that breaks the dependence between A and X.

For simplicity we introduce the method of instrumental variables in a linear setting. Suppose that (A, X, H, Y), with H unobserved, is generated by a linear SCM of the form

$$\begin{aligned} A &:= N_A, \quad H := N_H, \\ X &:= \xi_0^\intercal A + \delta_0^\intercal H + N_X, \\ Y &:= \gamma_0^\intercal X + \eta_0^\intercal H + N_Y \end{aligned}$$

for some mutually independent noise innovations N_A, N_H, N_X, N_Y and structural coefficients $\xi_0, \delta_0, \eta_0, \gamma_0 \neq 0$. The causal graph for this setup, corresponding to $A, H, X, Y \in \mathbb{R}$, is illustrated in Figure 1.8

Suppose, furthermore, that the covariance matrices $\mathrm{Var}(A)$ and $\mathrm{Var}(X)$ are positive definite. For notational simplicity, let $U := \eta_0^\intercal H + N_Y$ denote the unobserved variables entering the structural assignment of Y. Note that A satisfies the criteria of relevancy and exogeneity for being instruments for the causal effect from X to Y. The ordinary least squares method, in general, fails to be a consistent estimator of the causal effect γ from X to Y, i.e., the population OLS coefficient given by

$$\gamma_{\mathrm{OLS}} := \mathbb{E}[XX^\intercal]^{-1}\mathbb{E}[XY] = \gamma_0 + \mathbb{E}[XX^\intercal]^{-1}\mathbb{E}[XU^\intercal] \neq \gamma_0,$$

as $\mathbb{E}[XU] \neq 0$ due to the hidden confounding. On the other hand, if $\mathbb{E}[AX^\intercal]$ is of full column rank (known as the rank condition for identification which requires

that $|A| \geq |X|$), then we realize that the population two-stage least squares (TSLS) coefficient

$$\begin{aligned}\gamma_{\text{TSLS}} &:= (\mathbb{E}[XA^\intercal]\mathbb{E}[AA^\intercal]^{-1}E[AX^\intercal])^{-1}\mathbb{E}[XA^\intercal]\mathbb{E}[AA^\intercal]^{-1}E[AY] \\ &= \gamma_0 + (\mathbb{E}[XA^\intercal]\mathbb{E}[AA^\intercal]^{-1}E[AX^\intercal])^{-1}\mathbb{E}[XA^\intercal]\mathbb{E}[AA^\intercal]^{-1}E[AU] = \gamma_0,\end{aligned}$$

coincides with the causal effect from X to Y, as $E[AU] = 0$ by exogeneity. Thus, under the existence of instruments, the causal effect becomes identified from the observational distribution in the presence of hidden confounding. The name two-stage least squares come from the empirical counterpart to the population two-stage least squares coefficient coincides with the estimate resulting from a two-stage ordinary least squares procedure, where one first regresses X on A followed by a regression of Y on the first stage predicted values of X. The TSLS estimator can also be seen as a special case of the generalized method of moments (GMM), exploiting the moment restriction $\mathbb{E}[A(Y - \gamma^\intercal X)] = 0$ if and only if $\gamma = \gamma_0$ (see, e.g., Hall, 2005).

The instrumental variable method is also applicable in nonlinear structural causal models; In Chapter 3, we, for example, utilize that the existence of instruments can identify nonlinear causal functions. See Appendix B.2 for further discussion and references on nonlinear and nonparametric instrumental variable regression.

1.4.2.2. The P-Uncorrelated Least Squares Estimator

In Chapter 2, we propose a novel estimator in the linear instrumental variable setting called the p-uncorrelated least squares estimator (PULSE), which has the intuitive interpretation of minimizing the mean squared prediction error over a confidence region for the causal parameter. We show through simulation studies that our estimator, which can also be seen as a data-driven regularized TSLS regression, suffers from less variability than TSLS and other competing estimators while maintaining consistency. We continue our summary of the PULSE using the linear SCM setup of Section 1.4.2.1.

The two-stage least squares estimator is very unstable, especially in weak instrument settings (the effect from A to X is weak; see Appendix A.10 for further details). The TSLS estimator does not have moments of any order in the just-identified setup ($|A| = |X|$); see, e.g., Mariano (2001).

Under certain identifiability conditions, the null-hypothesis

$$\mathcal{H}_0(\alpha) : \text{Corr}(A, Y - X\alpha) = 0,$$

is only satisfied by the causal coefficient, i.e., the causal effect from X to Y. The TSLS estimator sets the sample covariance between instruments and the regression residuals to zero in the just-identified setup. Intuitively, this restriction might be too strong as the sample covariance, even for the true causal coefficients, is likely to be small but non-zero. On the other hand, the OLS estimate is known to be biased but fairly stable with moments of any order for sufficiently large sample

sizes (see, e.g., Mariano, 1972). The idea of the p-uncorrelated least squares estimator (PULSE) is to minimize the mean squared prediction error constrained to a finite-sample acceptance region $\mathcal{A}_n$ of a test for uncorrelatedness, $\mathcal{H}_0(\alpha)$. That is, we propose an estimator of the form

$$\hat{\gamma}^n_{\text{PULSE}} := \begin{array}{ll} \arg\min_\gamma & \frac{1}{n}\sum_{k=1}^n (Y_k - \gamma^\intercal X_k)^2 \\ \text{subject to} & \gamma \in \mathcal{A}_n. \end{array} \tag{1.1}$$

In Chapter 2, we propose a class of asymptotically valid hypothesis tests for $\mathcal{H}_0(\alpha)$. While the test has desirable properties the resulting minimization in Equation (1.1) becomes a non-convex optimization problem. However, through careful analysis and dual theory, we show that the estimator can be efficiently computed as

$$\hat{\gamma}^n_{\text{PULSE}} := l^n_{\text{OLS}}(\gamma) + \lambda^* l^n_{\text{IV}}(\gamma), \tag{1.2}$$

where l^n_{OLS} and $l^n_{\text{IV}}(\gamma)$ is the empirical ordinary and two-stage least squares loss functions (i.e., the OLS and TSLS estimators minimizes these functions, respectively) and λ^* is a data-dependent regularization parameter that can be approximated with arbitrary precision. This representation also reveals that the PULSE estimator belongs to a special class of estimators known as K-class estimators (Theil, 1953).

In an identified setup, the PULSE estimator consistently estimates the causal coefficient γ_0. In other words, the data-dependent λ^* is guaranteed to tend to infinity as the sample size increases. Hence, the data-driven mean squared prediction error (MSPE) regularization vanishes in the large sample limit. The PULSE estimator is also well-defined in the under-identified setup ($|A| < |X|$), which renders the causal effect unidentified. In the under-identified setup, the empirical objective is still to find the best predictive model among all coefficients that do not reject uncorrelatedness. Here, however, the target is not the causal coefficient but the coefficient in the TSLS solution space (all coefficients that render the instruments independent of residuals), which minimizes the MSPE.

Extensive simulation studies show that there are settings where the PULSE estimator indeed outperforms the TSLS and other competing instrumental variable estimators in terms of mean squared error (MSE). Weak instruments and weak endogeneity roughly characterize these settings. The MSPE regularization increases the bias in these settings, but the corresponding decrease in variance yields an MSE superior estimator. Furthermore, in Chapter 3, we extend this data-dependent MSPE regularization idea to nonlinear instrumental variable setups. The proposed estimator NILE likewise shows an MSE performance gain compared to various state-of-the-art nonparametric instrumental variable estimators.

1.5. Learning Generalizing Functions

Suppose that we are interested in learning prediction methods that minimize a particular loss function over the observational distribution. For example, it

is common to construct a prediction method that minimizes the mean squared prediction error (MSPE) over the observational distribution $\arg\min_{f_\diamond} \mathbb{E}[(Y - f_\diamond(X))^2]$, which we know coincides with the conditional expectation function of Y given X, but other loss functions may be reasonable too.

However, in many applications, one may wish to employ the prediction method on future system instances. For some systems, it may be reasonable to expect that future instances are subject to change. Alternatively, one may wish to employ a prediction method to entirely new systems known to differ from the system on which the method was trained. These problems are known under slight variations as, for example, covariate shift, domain generalization/adaption, and out-of-distribution generalization/prediction. We refer the reader to Section 3.1 of Chapter 3 for numerous references in this area of research. Common to these research areas is that the distribution of the training instance P_{train} differs from the class of possible test distributions $\mathcal{P}$ on which the prediction method is to be applied.

If one has a priori knowledge of the likelihood that each possible test distribution is to appear, one could, for example, try to minimize a weighted average of the MSPE over all possible test distributions. Alternatively, we may consider the problem of learning a prediction method f^* that seeks to minimize the worst-case MSPE;

$$f^* \in \arg\min_{f_\diamond} \sup_{P_{\text{test}} \in \mathcal{P}} \mathbb{E}_{P_{\text{test}}}[(Y - f_\diamond(X))^2],$$

where $\mathbb{E}_{P_{\text{test}}}$ denotes the expectation with respect to the distribution P_{test}. In this thesis, we concentrate on the latter objective. We say that a prediction method f^* is distributionally robust, a generalizing function, or a minimax solution with respect to a class of distributions $\mathcal{P}$ if it minimizes the worst-case prediction risk over all distributions in $\mathcal{P}$

In order to learn such a generalizing prediction method, we first must specify the class $\mathcal{P}$ of possible test distributions. A common approach is to say that the test distributions are slight variations of the training distribution in the sense that $P_{\text{test}} \in B_\rho(P_{\text{train}}, \varepsilon)$, i.e., that the test distribution lies within an ε-ball of the training distribution P_{train}, for some metric ρ on the space of probability measures, e.g., the Wasserstein metric. While this framework aims to guard against test distributions that arise from small perturbations in training distribution with respect to some probability metric, one may argue that it may be more natural for many applications that the test distributions arise from external manipulation of the original system.

In Chapter 3, we consider the problem of learning generalizing functions with respect to test distributions that are induced by interventions. That is, the set of possible test distributions $\mathcal{P} = \{P_{M(i)} : i \in \mathcal{I}\}$ are given by intervention-induced distributions in the underlying structural causal model M. Here, $\mathcal{I}$ denotes a class of interventions. We consider a framework where M belongs to a fairly general class of models $\mathcal{M}$ which both contains the response variable Y, predictors X,

latent variables H, and exogenous variables A. We allow for certain well-behaved interventions on X and A and aim to find generalizing prediction functions within some pre-specified function class $\mathcal{F}$. That is, we aim to learn

$$f^* \in \arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2],$$

where $\mathbb{E}_{M(i)}$ denotes the expectation with respect to the interventional distribution induced by the intervention i in the model M. Such generalizing prediction functions depend, among other things, on the function class $\mathcal{F}$, the model class $\mathcal{M}$ and the class of interventions $\mathcal{I}$.

It is well-known that when $\mathcal{I}$ contains all possible hard interventions of the form $\mathcal{I} = \{\text{do}(X := x) : x \in \mathbb{R}^d\}$ then the causal function f solves the minimax problem (see, e.g., Rojas-Carulla et al., 2018a). Conversely, we may also consider $\mathcal{I}$ to be a singleton consisting of the trivial intervention which does nothing, in which case $x \mapsto \mathbb{E}[Y|X = x]$ is a minimax solution.

We show, for example, that the causal function is a minimax solution even for singleton interventions that are confounding-removing, i.e., interventions that break the confounding between the predictors X and the target Y. Furthermore, we show that minimax solutions that differ from the causal function are highly susceptible to misspecifications of the intervention class. While the causal function is minimax whenever $\mathcal{I}$ contains at least one confounding-removing intervention, alternative non-causal minimax solutions may perform worse than the causal function if the intervention class is misspecified.

In practical scenarios, the underlying model M is unknown, and we do therefore not have access to the intervention induced-distributions $P_{M(i)}$ for $i \in \mathcal{I}$. Thus, similar to the hurdles plaguing the inference of causal effects from observational data, we can not identify and learn generalizing functions from observational data without further causal assumptions. There may exist an alternative model $\tilde{M} \in \mathcal{M}$ with identical observational distribution, $P_{\tilde{M}} = P_M$ but which differs on intervention distributions. As such, we say that distribution generalization is possible if there exists a function f^* which is minimax optimal for all observationally equivalent models within the model class $\mathcal{M}$.

We present sufficient conditions for distribution generalization in terms of restrictions on the observational distribution P_M, the intervention class $\mathcal{I}$, and the model class $\mathcal{M}$. Furthermore, we provide several impossibility theorems which illustrate the necessity of some of these restrictions.

1.5.1. PULSE and NILE

We know that when the intervention class contains arbitrarily strong interventions on X or at least one confounding-removing intervention then the causal function is a generalizing function. As such, any learning method for the causal function is equivalently learning a generalizing prediction function. The PULSE estimator, for example, consistently estimates a generalizing linear prediction function.

Similar considerations hold for nonlinear and nonparametric instrumental variable estimators as long as the intervention class is not support extending. That is, as long as the interventions do not extend the support of X. Since instrumental variable estimators can only recover the causal function on the support of the observational distribution this restriction is necessary without further causal assumptions. If, however, the interventions are support extending, then further causal assumptions are needed to extrapolate the estimate outside the support of X.

In Chapter 3, we present a nonlinear instrumental variable estimator which explicitly incorporates causal assumptions that the causal functions extrapolate linearly outside the support of observational distribution. We call this the nonlinear intervention-robust linear extrapolator (NILE). The linear extrapolation is not of importance — any extrapolation scheme which is uniquely determined by the on-support behavior works equally well. The NILE also uses the data-driven MSPE regularization ideas introduced for the PULSE.

1.5.2. Anchor Regression and K-class Estimators

For linear SCMs Rothenhäusler et al. (2021) show that among linear prediction functions, there exist functions that are minimax solutions but do not coincide with the causal functions whenever the intervention class $\mathcal{I}$ consists of bounded interventions on exogenous variables. That is, they show that for linear SCMs with exogenous variables A (called anchors), endogenous variables X, and a target Y, the anchor regression coefficient with regularization parameter λ is distributionally robust. More specifically, this linear prediction of Y from X is distributionally robust with respect to interventions on the exogenous variables A up to a certain strength that depends on λ.

The results of Chapter 2 shows that anchor regression is closely related to K-class estimators (Theil, 1953), which are parameterized by a real-valued parameter κ. The K-class estimators contain several well-known linear effect estimators: the ordinary least squares estimator for $\kappa = 0$, the two-stage least squares estimator for $\kappa = 1$, and for specific data-driven κ one can recover the limited information maximum likelihood (Anderson and Rubin, 1949) and Fuller estimators (Fuller, 1977).

Using the ideas of Rothenhäusler et al. (2021), we extend the distributional robustness property of anchor regression to general K-class estimators with fixed $\kappa \in [0, 1)$. Namely, we show that for a fixed $\kappa \in [0, 1)$ the K-class estimator $\hat{\alpha}^n_{\mathrm{K}}(\kappa, Y, Z, A)$ for regressing Y onto $Z \subseteq (X, A)$ using that A are exogenous variables, converges in probability towards a population quantity that is minimax prediction optimal among all linear predictors. That is,

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa, Y, Z, A) \xrightarrow{P}_n \arg\min_{\alpha} \sup_{v \in C(\kappa)} \mathbb{E}^{\mathrm{do}(A:=v)}[(Y - \alpha^\intercal Z)^2],$$

as the sample size n tends to infinity, where the intervention class is given by

$$C(\kappa) := \{v : \Omega \to \mathbb{R}^q : \mathbb{E}[vv^\intercal] \preceq (1-\kappa)^{-1}\mathbb{E}[AA^\intercal]\}.$$

Chapter 2

Distributional Robustness of K-class Estimators and the PULSE

Joint work with

Jonas Peters

Abstract

While causal models are robust in that they are prediction optimal under arbitrarily strong interventions, they may not be optimal when the interventions are bounded. We prove that the classical K-class estimator satisfies such optimality by establishing a connection between K-class estimators and anchor regression. This connection further motivates a novel estimator in instrumental variable settings that minimizes the mean squared prediction error subject to the constraint that the estimator lies in an asymptotically valid confidence region of the causal coefficient. We call this estimator PULSE (p-uncorrelated least squares estimator), relate it to work on invariance, show that it can be computed efficiently as a data-driven K-class estimator, even though the underlying optimization problem is non-convex, and prove consistency. We evaluate the estimators on real data and perform simulation experiments illustrating that PULSE suffers from less variability. There are several settings including weak instrument settings, where it outperforms other estimators.

Keywords: Causality, distributional robustness, instrumental variables

2.1. Introduction

Learning causal parameters from data has been a key challenge in many scientific fields and has been a long-studied problem in econometrics (e.g. Goldberger, 1972; Simon, 1953; Wold, 1954). Many years after the groundbreaking work by Fisher (1935) and Peirce (1883), causality plays again an increasingly important role in machine learning and statistics, two research areas that are most often considered part of mathematics or computer science (e.g., Imbens and Rubin, 2015; Pearl, 2009; Peters et al., 2017; Spirtes et al., 2000). Even though the current developments in

mathematics, computer science on the one and econometrics on the other hand do not forego independently, we believe that there is a lot of potential for more fruitful interaction between these two fields. Differences in the language have emerged, which can make communication difficult, but the target of inference, the underlying principles, and the methodology in both fields are closely related. This paper establishes a link between two developments in these fields: K-class estimation which aims at estimation of causal parameters with good statistical properties and invariance principles that are used to build methods that are robust with respect to distributional shifts. This connection allows us to prove distributional robustness guarantees for K-class estimators and motivates a new estimator, PULSE. We summarize our main results in Section 2.1.2.

2.1.1. Related Work

Given causal background knowledge, causal parameters can be estimated when taking into account confounding effects between treatment and outcome. Several related techniques have been suggested to tackle that problem, including variable adjustment (Pearl, 2009), propensity score matching (Rosenbaum and Rubin, 1983), inverse probability weighting (Horvitz and Thompson, 1952) or G-computation (Robins, 1986).

If some of the relevant variables have not been observed, one may instead use exogenous variation in the data to infer causal parameters, e.g., in the setting of instrumental variables (e.g., Imbens and Angrist, 1994; Newey, 2013; Wang and Tchetgen, 2018; Wright, 1928). Limited information estimators leverage instrumental variables to conduct single equation inference. An example of such methods is the two-stage least squares estimators (TSLS) developed by Theil (1953). Instead of minimizing the residual sum of squares as done by the ordinary least square (OLS) estimator, the TSLS minimizes the sample-covariance between the instruments and regression residuals. TSLS estimators are consistent, but are known to have suboptimal finite sample properties, e.g., they only have moments op to the degree of over-identification (Mariano, 1972). Kadane (1971) shows that under suitable conditions, the mean squared error of TSLS might even be larger than the one of OLS if the sample size is small (more precisely and using the notation introduced below, if $0 \leq n - q \leq 2(3 - (q_2 - d_1))$, where $q_2 - d_1$ is the degree of overidentification). This result is another indication that under certain conditions, it might be beneficial to use the OLS for regularization. Another method of inferring causal parameters in structural equation models is the limited information maximum likelihood (LIML) estimator due to Anderson and Rubin (1949). Theil (1958) introduced K-class estimators, which contain OLS, TSLS and the LIML estimator as special cases. This class of estimators is parametrized by a deterministic or stochastic parameter $\kappa \in [0, \infty)$ that depends on the observational data. Under mild regularity conditions a member of this class is consistent and asymptotically normally distributed if $(\kappa - 1)$ and $\sqrt{n}(\kappa - 1)$ converge, respectively, to zero in probability when n tends to infinity; see, e.g. Mariano (1975), Mariano

(2001). While the LIML does not have moments of any order, it shares the same asymptotic normal distribution with TSLS. Based on simulation studies, Anderson (1983) argues that, in many practically relevant cases, the normal approximation to a finite-sample estimator is inadequate for TSLS but a useful approximation in the case of LIML. Using Monte Carlo simulations, Hahn et al. (2004) recommend that the no-moment estimator LIML should not be used in weak instrument situations, where Fuller estimators have a substantially smaller MSE. The Fuller estimators (Fuller, 1977) form a subclass of the K-class estimators based on a modification to the LIML, which fixes the no-moment problem while maintaining consistency and asymptotic normality. Kiviet (2020) proposes a modification to the OLS estimator that makes use of explicit knowledge of the partial correlation between the covariates and the unobserved noise in Y. Andrews and Armstrong (2017) propose an unbiased estimator that is based on knowledge of the sign of the first stage regression and the variance the reduced form errors and that is less dispersed than TSLS, for example. Judge and Mittelhammer (2012) consider an affine combination of the OLS and TSLS estimators, which, again, yields a modification in the space of estimators. We prove that our proposed estimator, PULSE, can also be written as a data driven K-class estimator. As such, it minimizes a convex combination of the OLS and TSLS loss functions and can, in general, not be written as a convex combination of the estimators.

All of the above methods exploit background knowledge, e.g., in form of exogeneity of some of the variables. If no such background knowledge is available, it may still be possible, under additional assumptions, to infer the causal structure, e.g., represented by a graph, from observational (or observational and interventional) data. This problem is sometimes referred to as causal discovery. Constraint-based methods assume that the underlying distribution is Markov and faithful with respect to the causal graph and perform conditional independence tests to infer (parts of) the graph; see, e.g. Spirtes et al. (2000). Score-based methods assume a certain statistical model and optimize (penalized) likelihood scores; see, e.g. Chickering (2002). Some methods exploit a simple form of causal assignments, such as additive noise (e.g., Peters et al., 2014, and Shimizu et al., 2006) and others are based on exploiting invariance statements (e.g., Meinshausen et al., 2016; Peters et al., 2016). Many of such methods assume causal sufficiency, i.e., that all causally relevant variables have been observed, but some versions exist that allow for hidden variables; see, e.g. Claassen et al. (2013) and Spirtes et al. (1995).

Recent works in the fields of machine learning and computational statistics (e.g. Heinze-Deml and Meinshausen, 2021; Pfister et al., 2019; Schölkopf et al., 2012) investigate whether causal ideas can help to make machine learning methods more robust. The reasoning is that causal models are robust against any intervention in the following sense. Consider a target or response variable Y and covariates $X_1, \dots, X_p$. If we regress Y on the set X_S, $S \subseteq \{1, \dots, p\}$, of direct causes, then this regression function $x \mapsto E[Y|X_S = x]$ does not change when intervening on any of the covariates (which is sometimes referred to as 'invariance'). This statement

can be proved using the local Markov property (Lauritzen, 1996), for example, but the underlying fundamental principle has been discussed already several decades ago; most prominently using the terms 'autonomy' or 'modularity' (Haavelmo, 1944, and Aldrich, 1989). As a result, causal models of the form $x \mapsto E[Y|X_S = x]$ may perform well in prediction tasks, where, in the test distribution, the covariates have been intervened on. If, however, training and test distributions coincide, a model focusing only on prediction and the estimand $x \mapsto E[Y|X = x]$ may outperform a causal approach.

The two models described above (OLS and the causal model) formally solve a minimax problem on distributional robustness. Consider therefore an acyclic linear structural equation model (SEM) over (Y, X) with observational distribution F. Details on SEMs and interventions can be found in Appendix A.1. Assume that the assignment for Y equals $Y = \gamma_0^\intercal X + \varepsilon_Y$ for some $\gamma_0 \in \mathbb{R}^d$. The variables corresponding to non-zero entries in $\gamma_0^\intercal X$ are called the parents of Y, and ε_Y is assumed to be independent of these parents. Then, the mean squared prediction error when considering the observational distribution is not necessarily minimized by γ_0, that is, in general, we have $\gamma_0 \neq \gamma_{\text{OLS}} := \arg\min_\gamma E_F\left[(Y - \gamma^\intercal X)^2\right]$. Intuitively, we may improve the prediction of Y by including other variables than the parents of Y, such as its descendants. When considering distributional robustness, we are interested in finding a γ that minimizes the worst case expected squared prediction error over a class of distributions, $\mathcal{F}$, that is,

$$\arg\min_\gamma \sup_{F \in \mathcal{F}} E_F\left[(Y - \gamma^\intercal X)^2\right]. \tag{2.1}$$

If we observe data from all different distributions in $\mathcal{F}$ (and know which data point comes from which distribution), we can tackle this optimization directly (Meinshausen and Bühlmann, 2015). But estimators of Equation (2.1) may be available even if we do not observe data from each distribution in F. The true causal coefficient γ_0, for example, minimizes Equation (2.1) when $\mathcal{F}$ is the set of all possible (hard) interventions on X (e.g., Rojas-Carulla et al., 2018b). The OLS solution is optimal when $\mathcal{F}$ only contains the training distribution. In this sense, the OLS solution and the true causal coefficient constitutes the end points of a spectrum of estimators that are prediction optimal under a certain class of distributions.

Intuitively, models trading off causality and predictability may perform well in situations, where the test distribution is only moderately different from the training distribution. Anchor regression by Rothenhäusler et al. (2021), see Section 2.2.2 for details, is one approach formalizing this intuition in a linear setup. Similarly to an instrumental variable setting, one assumes the existence of exogenous variables that are called A (for anchor) which may or may not act directly on the target Y. The proposed estimator minimizes a convex combination of the residual sum of squares and the TSLS loss function and is shown to be prediction optimal in the sense of Equation (2.1) for a class $\mathcal{F}$ containing interventions on the covariates up to a certain strength; this strength depends on a regularization parameter:

the weight that is used in the convex combination of anchor regression. Other approaches (Magliacane et al., 2018; Pfister et al., 2021; Rojas-Carulla et al., 2018b) search over different subsets S and aim to choose sets that are both invariant and predictive.

2.1.2. Summary and Contributions

This paper contains two main contributions: A distributional robustness property of K-class estimators with fixed κ-parameter and a novel estimator for causal coefficients called the p-uncorrelated least squares estimator (PULSE). The following two sections summarize our contributions.

2.1.2.1. Distributional Robustness of K-class Estimators.

In Section 2.2 we show that anchor regression is closely related to K-class estimators. In particular, we prove that for a restricted subclass of models K-class estimators can be written as anchor regression estimators. For this subclass, this directly implies a distributional robustness property of K-class estimators. We then prove a similar robustness property for general K-class estimators with a fixed penalty parameter, and show that these properties hold even if the model is misspecified.

Consider a possibly cyclic linear SEM over the variables (Y, X, H, A) of the form

$$\begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} := \begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} B + A^\intercal M + \varepsilon^\intercal,$$

subject to regularity conditions that ensure the distribution of (Y, X, H, A) is well-defined. Here, B and M are constant matrices, the random vectors A and ε are defined on a common probability space $(\Omega, \mathcal{F}, P)$, Y is the endogenous target for the single equation inference, X are the observed endogenous variables, H are hidden endogenous variables and A are exogenous variables independent from the unobserved noise innovations ε.

SEMs allow for the notion of interventions, i.e., modeling external manipulations of the system. In this work, we are only concerned with interventions on the exogenous variables A of the form $\mathrm{do}(A := v)$. Because A is exogeneous, these interventions can be defined as follows: they change the distribution of A to that of a random vector v. The interventional distribution of the variables (Y, X, H, A) under the intervention $\mathrm{do}(A := v)$ is given by the simultaneous distribution of (X_v, Y_v, H_v, v) generated by the SEM

$$\begin{bmatrix} Y_v & X_v^\intercal & H_v^\intercal \end{bmatrix} := \begin{bmatrix} Y_v & X_v^\intercal & H_v^\intercal \end{bmatrix} B + v^\intercal M + \varepsilon.$$

Thus, the intervention does not change any of the original structural assignments of the endogenous variables. Instead, the change in the distribution of the exogeneous variable propagates through the system. We henceforth let $E^{\mathrm{do}(A:=v)}$ denote the expectation with respect to the interventional distribution of the system under the intervention $\mathrm{do}(A := v)$. More details on interventions can be found Appendix A.1

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{A})$ consist of n row-wise independent and identically distributed copies of the random vector (Y, X, H, A) and consider the single equation of interest

$$\mathbf{Y} = \mathbf{X}\gamma_0 + \mathbf{A}\beta_0 + \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y = \mathbf{X}\gamma_0 + \mathbf{A}\beta_0 + \tilde{\mathbf{U}}_Y.$$

The K-class estimator with parameter κ using non-sample information that only $\mathbf{Z}_* \subseteq [\mathbf{X}\ \mathbf{A}]$ have non-zero coefficients in the target equation of interest is given by

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa) = (\mathbf{Z}_*^\intercal(I - \kappa P^\perp_{\mathbf{A}})\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\intercal(I - \kappa P^\perp_{\mathbf{A}})\mathbf{Y},$$

where $P^\perp_{\mathbf{A}}$ is the projection onto the orthogonal complement of the column space of $\mathbf{A}$. For a fixed $\kappa \in [0, 1)$ K-class estimators can be represented by a penalized regression problem $\hat{\alpha}^n_{\mathrm{K}}(\kappa) = \arg\min_\alpha l^n_{\mathrm{OLS}}(\alpha) + \kappa/(1-\kappa) l^n_{\mathrm{IV}}(\alpha)$, where l^n_{OLS} and l^n_{IV} are the empirical OLS and TSLS loss functions, respectively. This representation and the ideas of Rothenhäusler et al. (2021) allow us to prove that K-class estimator converges to a coefficient that is minimax optimal when considering all distributions induced by a certain set of interventions of A. More specifically, we show that for a fixed κ and regardless of identifiability,

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa) \xrightarrow[n\to\infty]{P} \arg\min_\alpha \sup_{v \in C(\kappa)} E^{\mathrm{do}(A:=v)}\left[(Y - \alpha^\intercal Z_*)^2\right],$$

where $C(\kappa) := \{v : \Omega \to \mathbb{R}^q : \mathrm{Cov}(v, \varepsilon) = 0,\ E[vv^\intercal] \preceq \frac{1}{1-\kappa} E[AA^\intercal]\}$. The argmin on the right-hand side minimizes the worst case prediction error when considering interventions up to a certain strength (measured by the set $C(\kappa)$). This objective becomes relevant when we consider a response variable with several covariates and aim to minimize the mean squared prediction error of future realizations of the system of interest that do not follow the training distribution. The above result says that if the new realizations correspond to (unknown) interventions on the exogenous variables that are of bounded strength, K-class estimators with fixed $\kappa \in (0, 1)$ minimize the worst case prediction performance and, in particular, outperform the true causal parameter and the least squares solution (see also Figure A.2 in Section A.8.1). For κ approaching one, we recover the guarantee of the causal solution and for κ approaching zero, the set of distributions contains the training distribution. The above minimax property therefore adds to the discussion whether non-consistent K-class estimators with penalty parameter not converging to one can be useful; see, e.g. Dhrymes (1974).

2.1.2.2. The PULSE Estimator

Section 2.3 contains the second main contribution in this work. We propose a novel data driven K-class estimator for causal coefficients, which we call the p-uncorrelated least square estimator (PULSE). As above, we consider a single endogenous target in an SEM (or simultaneous equation model) and aim to predict it from observed predictors that are with a priori (non-sample) information known

to be either endogenous or exogenous. The PULSE estimator can be written in several equivalent forms. It can, first, be seen as a data-driven K-class estimator

$$\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n/(1+\lambda^\star_n)) = \arg\min_\alpha l^n_{\mathrm{OLS}}(\alpha) + \lambda^\star_n l^n_{\mathrm{IV}}(\alpha),$$

where

$$\lambda^\star_n := \inf\left\{\lambda > 0 \ : \ \begin{array}{c}\text{testing } \mathrm{Corr}(A, Y - Z\hat{\alpha}^n_{\mathrm{K}}(\lambda/(1+\lambda))) = 0 \\ \text{yields a p-value} \geq p_{\min}\end{array}\right\},$$

for some pre-specified level of the hypothesis test $p_{\min} \in (0,1)$. In words, the PULSE estimator outputs the K-class estimator closest to the OLS while maintaining a non-rejected test of uncorrelatedness. In principle, PULSE can be used with any testing procedure. The choice of test, however, may influence the difficulty of the resulting optimization problem. In this paper, we investigate PULSE in connection with a specific class of hypothesis tests that, for example, contain the test of Anderson and Rubin (1949). For these hypothesis tests we develop an efficient and provably correct optimization method, that is based on binary line search and quadratic programming.

We show that our estimator can, second, be written as the solution to a constrained optimization problem. To that end, define the primal problems

$$\hat{\alpha}^n_{\mathrm{Pr}}(t) := \begin{array}{ll}\arg\min_\alpha & l^n_{\mathrm{OLS}}(\alpha) \\ \text{subject to} & l^n_{\mathrm{IV}}(\alpha) \leq t.\end{array}$$

For the choice $t^\star_n := \sup\{t \ : \ \text{testing } \mathrm{Corr}(A, Y - Z\hat{\alpha}^n_{\mathrm{Pr}}(t)) = 0 \text{ yields a } p\text{-value} \geq p_{\min}\}$, we provide a detailed analysis proving that $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n/(1+\lambda^\star_n)) = \hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n)$.

For the testing procedure proposed in this paper, we show that, third, PULSE can be written as

$$\begin{array}{ll}\mathrm{argmin}_\alpha & l^n_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}) \\ \text{subject to} & \alpha \in \mathcal{A}_n(1 - p_{\min}),\end{array}$$

where $\mathcal{A}_n(1-p_{\min})$ is the non-convex acceptance region for our test of uncorrelatedness.

This third formulation allows for a simple interpretation of our estimator: among all coefficients (not restricted to K-class estimators) that do not yield a rejection of uncorrelatedness, we choose the one that yields the best prediction. If the acceptance region is empty it outputs a warning indicating a possible model misspecification or an assumption violation to the user (in that case, one can formally output another estimator such as TSLS or Fuller, yielding PULSE well-defined).

In the just-identified setup, the TSLS estimator solves a normal equation which is equivalent to setting a sample covariance between the instruments and the resulting prediction residuals to zero; it then corresponds to $t = 0$. For this (and the over-identified) setting, we prove that PULSE is a consistent estimator for the causal coefficient.

The TSLS does not have a finite variance if there is insufficient degree of overidentification, for example. In particular for weak instruments, this usually comes with poor finite sample performance. In such cases, however, the acceptance region of uncorrelatedness is usually large. This yields a weak constraint in the optimization problem and the PULSE will be closer to the OLS, which in certain settings suffers from less variability (see, e.g., Hahn and Hausman, 2005; Hahn et al., 2004). In simulations we indeed see that, similarly to other data-driven K-class estimators that are pulled towards the OLS, such as Fuller estimators, the PULSE comes with beneficial finite sample properties compared to TSLS and LIML.

Unlike other estimators such as LIML or the classical TSLS, the PULSE is well-defined in under-identified settings, too. Here, its objective is still to find the best predictive solution among all parameters that do not reject uncorrelatedness. Uncorrelatedness to the exogeneous variable is sometimes referred to as invariance. The idea of choosing the best predictive among all invariant models has been investigated in several works (e.g. Magliacane et al., 2018; Pfister et al., 2021; Rojas-Carulla et al., 2018b) with the motivation to find models that generalize well (in particular, with respect to interventions on the exogenous variables). Existing methods, however, focus on selecting subsets of variables and then consider least squares regression of the response variable onto the full subset. PULSE can recover such type of solutions if they are indeed optimal. But it also allows to search over coefficients that are different from least squares regression for sets of variables. Consequently, PULSE allows us to find solutions in situations, where the above methods would not find any invariant subsets, which may often be the case if there are hidden variables (see Section A.8.3 for an example).

We show in a simulation study that there are several settings in which PULSE outperforms existing estimators both in terms of MSE ordering and several one-dimensional scalarizations of the MSE. More specifically, we show that PULSE can outperform the TSLS and Fuller estimators in weak instrument situations, for example, where Fuller estimators are known to have good MSE properties; see, e.g. Hahn et al. (2004) and Stock et al. (2002).

Implementation of PULSE and code for experiments (R) are available on GitHub.[1]

2.2. Robustness Properties of K-class Estimators

In this section we consider K-class estimators (Theil, 1958, and Nagar, 1959) and show a connection with anchor regression of Rothenhäusler et al. (2021). In Section 2.2.3.1 we establish the connection in models where we use *a priori* information that there are no included exogenous variables in the target equation of interest. In Section 2.2.3.2 we then show that general K-class estimators can be written as the solution to a penalized regression problem. In Section 2.2.3.3 we

[1] `https://github.com/MartinEmilJakobsen/PULSE`

utilize this representation and the ideas of Rothenhäusler et al. (2021) to prove a distributional robustness guarantee of general K-class estimators with fixed $\kappa \in [0, 1)$, even under model misspecification and non-identifiability. Proofs of results in this section can be found in Appendix A.3.

2.2.1. Setup and Assumptions

Denote the random vectors $Y \in \mathbb{R}, X \in \mathbb{R}^d, A \in \mathbb{R}^q, H \in \mathbb{R}^r$ and $\varepsilon \in \mathbb{R}^{d+1+r}$ by the target, endogenous regressor, anchors, hidden and noise variables, respectively. Let further (Y, X, H) be generated by the possibly cyclic structural equation model (SEM)

$$\begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} := \begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} B + A^\intercal M + \varepsilon^\intercal, \tag{2.1}$$

for some random vectors $\varepsilon \perp\!\!\!\perp A$ and constant matrices B and M. Let $(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{A})$ consist of $n \geq \min\{d, q\}$ row-wise independent and identically distributed copies of the random vector (Y, X, H, A). Solving for the endogenous variables we get the structural and reduced form equations $[\,\mathbf{Y}\ \mathbf{X}\ \mathbf{H}\,]\,\Gamma = \mathbf{A}M + \boldsymbol{\varepsilon}$ and $[\,\mathbf{Y}\ \mathbf{X}\ \mathbf{H}\,] = \mathbf{A}\Pi + \boldsymbol{\varepsilon}\Gamma^{-1}$, where $\Gamma := I - B$ and $\Pi := M\Gamma^{-1}$. Assume without loss of generality that Γ has a unity diagonal, such that the target equation of interest is given by

$$\mathbf{Y} = \mathbf{X}\gamma_0 + \mathbf{A}\beta_0 + \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y = \mathbf{Z}\alpha_0 + \tilde{\mathbf{U}}_Y, \tag{2.2}$$

where $(1, -\gamma_0, -\eta_0) \in \mathbb{R}^{(1+d+r)}$, $\beta_0 \in \mathbb{R}^q$ and $\boldsymbol{\varepsilon}_Y$ are the first columns of Γ, M and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ respectively, $\mathbf{Z} := [\mathbf{X}\ \mathbf{A}]$, $\alpha_0 = (\gamma_0, \beta_0) \in \mathbb{R}^{d+q}$ and $\tilde{\mathbf{U}}_Y := \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y$.

The possible dependence between the noise $\tilde{\mathbf{U}}_Y$ and the endogenous variables, i.e., the influence by hidden variables, generally, renders the standard OLS approach for estimating α_0 inconsistent. Instead, one can make use of the components in A that have vanishing coefficient in Equation (2.2) for consistent estimation. In the remainder of this work, we disregard any *a priori* (non-sample) information not concerning the target equation. The question of identifiability of α_0 has been studied extensively (Frisch, 1938; Haavelmo, 1944; Koopmans et al., 1950) and more recent overviews can be found in, e.g., Didelez et al. (2010), Fisher (1966), and Greene (2003).

We will use the following assumptions concerning the structure of the SEM:

Assumption 2.1 (Global assumptions)**.** *(a) (Y, X, H, A) is generated in accordance with the SEM in Equation (2.1); (b) $\rho(B) < 1$ where $\rho(B)$ is the spectral radius of B; (c) ε has jointly independent marginals $\varepsilon_1, \ldots, \varepsilon_{d+1+r}$; (d) A and ε are independent; (e) No variable in Y, X and H is an ancestor of A, that is, A is exogenous; (f) $E[\|\varepsilon\|_2^2]$, $E[\|A\|_2^2] < \infty$; (g) $E[\varepsilon] = 0$. (h) $Var(A) \succ 0$, i.e., the variance matrix of A is positive definite; (i) $\mathbf{A}^\intercal\mathbf{A}$ is almost surely of full rank;*

Assumption 2.2 (Finite sample assumptions)**.** *(a) $\mathbf{Z}_*^\intercal\mathbf{Z}_*$ is almost surely of full rank; (b) $\mathbf{A}^\intercal\mathbf{Z}_*$ is almost surely of full column rank. (c) $\mathbf{X}^\intercal\mathbf{X}$ is almost surely of full rank;*

Assumption 2.3 (Population assumptions)**.** *(a) $Var(Z_*) \succ 0$, i.e., the variance matrix of Z_* is positive definite; (b) $E[AZ_*^\intercal]$ is of full column rank.*

We will henceforth assume that Assumption 2.1 always holds. This assumption ensure that the SEM and that the TSLS objectives are well-defined. In the above assumptions, Z_* and $\mathbf{Z}_*$ are generic placeholders for a subset of endogenous and exogenous variables from $[X^\intercal\ A^\intercal]^\intercal$ and $[\mathbf{X}\ \mathbf{A}]$, respectively, which should be clear from the context in which they are used. Both Assumption 2.1.(i) and Assumption 2.2.(c) hold if X and A have density with respect to Lebesgue measure, which in turn is guaranteed by Assumption 2.1.(d) if A and ε have density with respect to Lebesgue measure. Assumption 2.1.(h) and 2.1.(i) implies that the instrumental variable objective functions introduced below is almost surely well-defined and Assumption 2.2.(c) yields that the ordinary least square solution is almost surely well-defined. Assumption 2.1.(f) implies that Y, X and H all have finite second moments. For Assumption 2.3.(b) and 2.2.(b) it is necessary that $q \geq \dim(Z_*)$, i.e., that the setup must be just- or over-identified; see Section 2.3.1 below.

2.2.2. Distributional Robustness of Anchor Regression

Rothenhäusler et al. (2021) proposes a method, called anchor regression, for predicting the endogenous target variable Y from the endogenous variables X. The collection of exogenous variables A, called anchors, are not included in that prediction model. Anchor regression trades off predictability and invariance by considering a convex combination of the ordinary least square (OLS) loss function and the two-stage least square (IV) loss function using the anchors as instruments. More formally, we define

$$l_{\mathrm{OLS}}(\gamma; Y, X) := E(Y - \gamma^\intercal X)^2, \tag{2.3}$$
$$l_{\mathrm{IV}}(\gamma; Y, X, A) := E(A(Y - \gamma^\intercal X))^\intercal E(AA^\intercal)^{-1} E(A(Y - \gamma^\intercal X)),$$
$$l^n_{\mathrm{OLS}}(\gamma; \mathbf{Y}, \mathbf{X}) := n^{-1}(\mathbf{Y} - \mathbf{X}\gamma)^\intercal(\mathbf{Y} - \mathbf{X}\gamma), \tag{2.4}$$
$$l^n_{\mathrm{IV}}(\gamma; \mathbf{Y}, \mathbf{X}, \mathbf{A}) := n^{-1}(\mathbf{Y} - \mathbf{X}\gamma)^\intercal P_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}\gamma), \tag{2.5}$$

the population and finite sample versions of the loss functions. Here $P_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal$ is the orthogonal projection onto the column space of $\mathbf{A}$. To simplify notation, we omit the dependence on Y, X, A, $\mathbf{A}$, $\mathbf{X}$ or $\mathbf{Y}$ when they are clear from a given context. For a penalty parameter $\lambda > -1$, the anchor regression coefficients are defined as

$$\gamma_{\mathrm{AR}}(\lambda) := \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^d} \{l_{\mathrm{OLS}}(\gamma) + \lambda l_{\mathrm{IV}}(\gamma)\},$$
$$\hat{\gamma}^n_{\mathrm{AR}}(\lambda) := \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^d} \{l^n_{\mathrm{OLS}}(\gamma) + \lambda l^n_{\mathrm{IV}}(\gamma)\}. \tag{2.6}$$

The estimator $\hat{\gamma}^n_{\mathrm{AR}}(\lambda)$ consistently estimates the population estimand $\gamma_{\mathrm{AR}}(\lambda)$ and minimizes prediction error while simultaneously penalizing a transformed sample

covariance between the anchors and the resulting prediction residuals. Unlike the TSLS estimator, for example, the anchor regression estimator is almost surely well-defined under the rank condition of Assumption 2.2.(c), even if the model is under-identified, that is, there are less exogenous than endogenous variables. The solution to the empirical minimization problem of anchor regression is given by

$$\hat{\gamma}^n_{\mathrm{AR}}(\lambda) = [\mathbf{X}^\intercal(I + \lambda P_{\mathbf{A}})\mathbf{X}]^{-1}\mathbf{X}^\intercal(I + \lambda P_{\mathbf{A}})\mathbf{Y}, \tag{2.7}$$

which follows from solving the normal equation of Equation (2.6).

The motivation of anchor regression is not to infer a causal parameter. Instead, for a fixed penalty parameter λ, the estimator is shown to possess a distributional or interventional robustness property: the estimator is optimal when predicting under interventions on the exogenous variables that are below a certain intervention strength. By Theorem 1 of Rothenhäusler et al. (2021) it holds that

$$\gamma_{\mathrm{AR}}(\lambda) = \operatorname*{arg\,min}_{\gamma\in\mathbb{R}^d} \sup_{v\in C(\lambda)} E^{\mathrm{do}(A:=v)}\left[(Y - \gamma^\intercal X)^2\right],$$

where $C(\lambda) := \{v : \Omega \to \mathbb{R}^q : \mathrm{Cov}(v, \varepsilon) = 0, E(vv^\intercal) \preceq (\lambda + 1)E(AA^\intercal)\}$.

2.2.3. Distributional Robustness of K-class Estimators

We now introduce the limited information estimators known as K-class estimators (Theil, 1958, and Nagar, 1959) used for single equation inference. Suppose that we are given non-sample information about which components of γ_0 and β_0, of Equation (2.2), are zero. We can then partition $\mathbf{X} = [\mathbf{X}_* \ \mathbf{X}_{-*}] \in \mathbb{R}^{n\times(d_1+d_2)}$, $\mathbf{A} = [\mathbf{A}_* \ \mathbf{A}_{-*}] \in \mathbb{R}^{n\times(q_1+q_2)}$ and $\mathbf{Z} = [\mathbf{Z}_* \ \mathbf{Z}_{-*}] = [\mathbf{X}_* \ \mathbf{A}_* \ \mathbf{X}_{-*} \ \mathbf{A}_{-*}]$ with $\mathbf{Z} \in \mathbb{R}^{n\times((d_1+q_1)+(d_2+q_2))}$, where $\mathbf{X}_{-*}$ and $\mathbf{A}_{-*}$ corresponds to the variables for which our non-sample information states that the components of γ_0 and β_0 are zero, respectively. We call the variables corresponding to $\mathbf{A}_*$ included exogenous variables. Similarly, we write $\gamma_0 = (\gamma_{0,*}, \gamma_{0,-*})$, $\beta_0 = (\beta_{0,*}, \beta_{0,-*})$ and $\alpha_0 = (\alpha_{0,*}, \alpha_{0,-*}) = (\gamma_{0,*}, \beta_{0,*}, \gamma_{0,-*}, \beta_{0,-*})$. The structural equation of interest then reduces to $\mathbf{Y} = \mathbf{X}_*\gamma_{0,*} + \mathbf{X}_{-*}\gamma_{0,-*} + \mathbf{A}_*\beta_{0,*} + \mathbf{A}_{-*}\beta_{0,-*} + \tilde{\mathbf{U}}_Y = \mathbf{Z}_*\alpha_{0,*} + \mathbf{U}_Y$, where $\mathbf{U}_Y = \mathbf{X}_{-*}\gamma_{0,-*} + \mathbf{A}_{-*}\beta_{0,-*} + \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y$. In the case that the non-sample information is indeed correct, we have that $\mathbf{U}_Y = \tilde{\mathbf{U}}_Y = \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y$. When well-defined, the K-class estimator with parameter $\kappa \in \mathbb{R}$ for a simultaneous estimation of $\alpha_{0,*}$ is given by

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) = (\mathbf{Z}_*^\intercal(I - \kappa P^\perp_{\mathbf{A}})\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\intercal(I - \kappa P^\perp_{\mathbf{A}})\mathbf{Y}, \tag{2.8}$$

where $I - \kappa P^\perp_{\mathbf{A}} = I - \kappa(I - P_{\mathbf{A}}) = (1 - \kappa)I + \kappa P_{\mathbf{A}}$.

Comparing Equations (2.7) and (2.8) suggests a close connection between anchor regression and K-class estimators for inference of structural equations with no included exogenous variables. In the following subsections, we establish this connection and subsequently extend the distributional robustness property to general K-class estimators.

2.2.3.1. K-class Estimators in Models with no Included Exogenous Variables

Assume that, in addition to Assumption 2.1, we have the non-sample information that $\beta_0 = 0$, that is, no exogenous variable in A directly affects the target variable Y. By direct comparison we see that the K-class estimator for $\kappa < 1$ coincides with the anchor regression estimator with penalty parameter $\lambda = \kappa/(1-\kappa)$, i.e., $\hat{\gamma}^n_{\mathrm{K}}(\kappa) = \gamma^n_{\mathrm{AR}}\left(\frac{\kappa}{1-\kappa}\right)$. Equivalently, we have $\gamma^n_{\mathrm{AR}}(\lambda) = \gamma^n_{\mathrm{K}}(\lambda/(1+\lambda))$ for any $\lambda > -1$. As such, the K-class estimator, for a fixed κ, inherits the following distributional robustness property:

$$\gamma_{\mathrm{K}}(\kappa) = \gamma_{\mathrm{AR}}\left(\frac{\kappa}{1-\kappa}\right) = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^d} \sup_{v \in C(\kappa/(1-\kappa))} E^{\mathrm{do}(A:=v)}\left[(Y - \gamma^\intercal X)^2\right], \tag{2.9}$$

where $C(\kappa/(1-\kappa)) = \{v : \Omega \to \mathbb{R}^q : \mathrm{Cov}(v, \varepsilon) = 0, E[vv^\intercal] \preceq \frac{1}{1-\kappa} E[AA^\intercal]\}$. This statement holds by Theorem 1 of Rothenhäusler et al. (2021).

In an identifiable model with $P\lim_{n\to\infty} \kappa = 1$ we have that $\hat{\gamma}^n_{\mathrm{K}}(\kappa)$ consistently estimates the causal parameter; see e.g. Mariano (2001). For such a choice of κ, the robustness above is just a weaker version of what the causal coefficient can guarantee. However, the above result in Equation (2.9) establishes a robustness property for fixed $\kappa < 1$, even in cases where the model is not identifiable. Furthermore, since we did not use that the non-sample information that $\beta_0 = 0$ was true, the robustness property is resilient to model misspecification in terms of excluding included exogenous variables from the target equation which generally also breaks identifiability.

2.2.3.2. The K-class Estimators as Penalized Regression Estimators

We now show that general K-class estimators can be written as solutions to penalized regression problems. The first appearance of such a representation is, to the best of our knowledge, due to McDonald (1977) building upon previous work of Basmann (1960a,b). Their representation, however, concerns only the endogenous part γ. We require a slightly different statement and will show that the entire K-class estimator of $\alpha_{0,*}$, i.e., the simultaneous estimation of $\gamma_{0,*}$ and $\beta_{0,*}$, can be written as a penalized regression problem. Let therefore $l_{\mathrm{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A})$, $l^n_{\mathrm{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A})$ and $l_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}_*)$, $l^n_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}_*)$ denote the population and empirical TSLS and OLS loss functions as defined in Equations (2.3) to (2.4). That is, the TSLS loss function for regressing $\mathbf{Y}$ on the included endogenous and exogenous variables $\mathbf{Z}_*$ using the exogeneity of $\mathbf{A}$ and $\mathbf{A}_{-*}$ as instruments and the OLS loss function for regressing $\mathbf{Y}$ on $\mathbf{Z}_*$. We define the K-class population and finite-sample loss functions as an affine combination of the two loss functions above. That is,

$$l_{\mathrm{K}}(\alpha; \kappa, Y, Z_*, A) = (1-\kappa) l_{\mathrm{OLS}}(\alpha; Y, Z_*) + \kappa l_{\mathrm{IV}}(\alpha; Y, Z_*, A), \tag{2.10}$$

$$l^n_{\mathrm{K}}(\alpha; \kappa, \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) = (1-\kappa) l^n_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}_*) + \kappa l^n_{\mathrm{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}). \tag{2.11}$$

Proposition 2.1. *Consider one of the following scenarios: 1) $\kappa < 1$ and 2.2.(a) holds, or 2) $\kappa = 1$ and 2.2.(b) holds. The estimator minimizing the empirical loss function of Equation (2.11) is almost surely well-defined and coincides with the K-class estimator of Equation (2.8). That is, it almost surely holds that*

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) = \underset{\alpha \in \mathbb{R}^{d_1+q_1}}{\arg\min}\, l^n_{\mathrm{K}}(\alpha; \kappa, \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}). \tag{2.12}$$

Assuming $\kappa \neq 1$, we can rewrite Equation (2.12) to

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) = \underset{\alpha \in \mathbb{R}^{d_1+q_1}}{\arg\min} \{ l^n_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}_*) + \frac{\kappa}{1-\kappa} l^n_{\mathrm{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) \}. \tag{2.13}$$

Thus, K-class estimators seek to minimize the ordinary least squares loss for regressing $\mathbf{Y}$ on $\mathbf{Z}_*$, while simultaneously penalizing the strength of a transform on the sample covariance between the prediction residuals and collection of exogenous variables $\mathbf{A}$.

In the following section, we consider a population version of the above quantity. If we replace the finite sample Assumption 2.2 with the corresponding population Assumption 2.3, we get that the minimization estimator of the empirical loss function of Equation (2.11) is asymptotically well-defined. Furthermore, we now prove that whenever the population assumptions are satisfied, then, for any fixed $\kappa \in [0, 1]$, $\hat{\alpha}^n_{\mathrm{K}}(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A})$ converges in probability towards the population K-class estimand.

Proposition 2.2. *Consider one of the following scenarios: 1) $\kappa \in [0, 1)$ and Assumption 2.3.(a) holds, or 2) $\kappa = 1$ and Assumption 2.3.(b) holds. It holds that $(\hat{\alpha}^n_K(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}))_{n \geq 1}$ is an asymptotically well-defined sequence of estimators. Furthermore, the sequence consistently estimates the well-defined population K-class estimand. That is,*

$$\hat{\alpha}^n_{\mathrm{K}}(\kappa; \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) \xrightarrow[n \to \infty]{P} \alpha_{\mathrm{K}}(\kappa; Y, Z_*, A) := \underset{\alpha \in \mathbb{R}^{d_1+q_1}}{\arg\min}\, l_{\mathrm{K}}(\alpha; \kappa, Y, Z_*, A).$$

2.2.3.3. Distributional Robustness of General K-class Estimators

We are now able to prove that the general K-class estimator possesses a robustness property similar to the statements above. It is prediction optimal under a set of interventions, now including interventions on all exogenous A up to a certain strength.

Theorem 2.1. *Let Assumption 2.1 hold. For any fixed $\kappa \in [0, 1)$ and $Z_* = (X_*, A_*)$ with $X_* \subseteq X$ and $A_* \subseteq A$, we have, whenever the population K-class estimand is well-defined, that*

$$\alpha_{\mathrm{K}}(\kappa; Y, Z_*, A) = \underset{\alpha \in \mathbb{R}^{d_1+q_1}}{\arg\min} \sup_{v \in C(\kappa)} E^{\mathrm{do}(A:=v)} \left[(Y - \alpha^\intercal Z_*)^2 \right],$$

where $C(\kappa) := \left\{ v : \Omega \to \mathbb{R}^q : \mathrm{Cov}(v, \varepsilon) = 0,\ E[vv^\intercal] \preceq \frac{1}{1-\kappa} E[AA^\intercal] \right\}$.

Here, $E^{\mathrm{do}(A:=v)}$ denotes the expectation with respect to the distribution entailed under the intervention $\mathrm{do}(A := v)$ (see Section 2.1.2.1 and Appendix A.1) and $(\Omega, \mathcal{F}, P)$ is the common background probability space on which A and ε are defined.

In words, among all linear prediction methods of Y using Z_* as predictors, the K-class estimator with parameter κ has the lowest possible worst case mean squared prediction error when considering all interventions on the exogenous variables A contained in $C(\kappa)$. As κ approaches one, the estimator is prediction optimal under a class of arbitrarily strong interventions in the direction of the variance of A. (Here, κ is arbitrary but fixed; the statement does not cover data-driven choices of κ, such as LIML or Fuller.) The above result is a consequence of the relation between anchor regression and K-class estimators. The special case $A_* = \emptyset$ is a consequence of Theorem 1 by Rothenhäusler et al. (2021). Our proof follows similar arguments but additionally allows for $A_* \neq \emptyset$.

The property in Theorem 2.1 has a decision-theoretic interpretation (see Chamberlain (2007) for an application of decision theory in IV models based on another loss function). Consider a response Y, covariates Z_* and a distribution (specified by θ) over (Y, Z_*), and the squared loss $\ell(Y, Z, \alpha) := (Y - \alpha^\top Z_*)^2$. Then, assuming finite variances, for each distribution the risk $E_\theta[(Y - \alpha^\top Z_*)^2]$ is minimized by the (population) OLS solution $\alpha = \alpha_\theta := \mathrm{cov}_\theta(Z_*)^{-1}\mathrm{cov}_\theta(Z_*, Y)$. In the setting of Theorem 2.1, we are given a distribution over (Y, Z_*), specified by θ, but we are interested in minimizing the risk $E_{\theta,v}[(Y - \alpha_\theta^\top Z_*)^2]$ for another distribution that is induced by an intervention and specified by (θ, v). The above result states that the K-class estimator minimizes a worst-case risk when considering all $v \in C(\kappa)$.

Theorem 2.1 makes use of the language of SEMs in that it yields the notion of interventions.[2] As such, the result can be rephrased using other causal frameworks. The crucial assumptions are the exogeneity of A and the linearity of the system. Furthermore, the result is robust with respect to several types of model misspecifications that breaks identifiability of α_0, such as excluding included endogenous or exogenous predictors or the existence of latent variables; see Remark A.1 in Appendix A.7.

2.3. The P-Uncorrelated Least Square Estimator

We now introduce the p-uncorrelated least square estimator (PULSE). As discussed in Section 2.1.2, PULSE allows for different representations. In this section we start with the third representation and show the equivalence of the other representations afterwards.

Consider predicting the target Y from endogenous and possibly exogenous regressors Z. Let therefore $\mathcal{H}_0(\alpha)$ denote the hypothesis that the prediction residuals using α as a regression coefficient is simultaneously uncorrelated with every exogenous variable, that is, $\mathcal{H}_0(\alpha) : \mathrm{Corr}(A, Y - \alpha^\intercal Z) = 0$. This hypothesis is

[2] In particular, we have not considered the SEM as a model for counterfactual statements.

in some models under certain conditions equivalent to the hypothesis that α is the true causal coefficient. One of these conditions is the rank condition Assumption 2.8 introduced below, also known as the rank condition for identification; Wooldridge (2010).

The two-stage least square (TSLS) estimator exploits the equivalence between the causal coefficient and the zero correlation between the instruments and the regression residuals. Here, one minimizes a sample covariance between the instruments and the regression residuals: we can write $l^n_{\text{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A}) = \|\widehat{\text{Cov}}_n(A, Y - \alpha^\intercal Z)\|^2_{(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1}}$ when A is mean zero.[3] In the just-identified setup the TSLS estimator yields a sample covariance that is exactly zero and is known to be unstable, in that it has no moments of any order. Intuitively, the constraint of vanishing sample covariance may be too strong.

Let $T(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A})$ be a finite sample test statistic for testing the hypothesis $\mathcal{H}_0(\alpha)$ and let p-value$(T(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A}))$ denote the p-value associated with the test of $\mathcal{H}_0(\alpha)$. We then define the p-uncorrelated least square estimator (PULSE) as

$$\hat{\alpha}^n_{\text{PULSE}}(p_{\min}) = \begin{array}{ll} \text{argmin}_\alpha & l^n_{\text{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}) \\ \text{subject to} & \text{p-value}(T(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A})) \geq p_{\min}, \end{array} \tag{2.1}$$

where $p_{\min}$ is a pre-specified level of the hypothesis test. In words, we aim to minimize the mean squared prediction error among all coefficients which yield a p-value for testing $\mathcal{H}_0(\alpha)$ that does not fall below some pre-specified level-threshold $p_{\min} \in (0, 1)$, such as $p_{\min} = 0.05$. That is, the minimization is constrained to the acceptance region of the test, i.e., a confidence region for the causal coefficient in the identified setup. Among these coefficient, we choose the solution that is 'closest' to the OLS solution.[4]

Thus, PULSE allows for an intuitive interpretation. We will see in the experimental section that it has good finite sample performance, in particular for weak instruments. Unlike other estimators, such as LIML, the above estimator is well-defined in the under-identified setup, too.[5] In such cases, PULSE extends on existing literature that aims to trade-off predictability and invariance but that so far has been restricted to search over subsets of variables (see Section 2.1.2.2 and Section A.8.3). To maintain consistency of the estimator the chosen test must have asymptotic power of one.

In this paper, we propose a class of significance tests, that contains, e.g., the Anderson-Rubin test (Anderson and Rubin, 1949). While the objective function in

[3] $\|\cdot\|_{(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1}}$ is the norm induced by the inner product $\langle x, y\rangle = x^\intercal(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1}y$.

[4] Here, closeness is measured in the OLS distance: We define the OLS norm via $\|\alpha\|^2_{\text{OLS}} := l^n_{\text{OLS}}(\alpha + \hat{\alpha}^n_{\text{OLS}}) - l^n_{\text{OLS}}(\hat{\alpha}^n_{\text{OLS}}) = \alpha^\top \mathbf{Z}^T\mathbf{Z}\alpha$, where $\hat{\alpha}^n_{\text{OLS}}$ is the OLS estimator. This defines a norm (rather than a semi-norm) if $\mathbf{Z}^T\mathbf{Z}$ is non-degenerate. Minimizing $l^n_{\text{OLS}}(\alpha) = \|\mathbf{Y} - \mathbf{Z}\alpha\|^2_2 = (\alpha - \hat{\alpha}^n_{\text{OLS}})^\intercal\mathbf{Z}^\intercal\mathbf{Z}(\alpha - \hat{\alpha}^n_{\text{OLS}}) + \|\mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\text{OLS}}\|^2_2$ is equivalent to minimizing $\|\alpha - \hat{\alpha}^n_{\text{OLS}}\|^2_{\text{OLS}}$.

[5] The PULSE estimator is defined for finite samples, but the following deliberation may help to build intuition: In an under-identified IV setting, minimizing $l_{\text{OLS}}(\gamma)$ under the constraint that $l_{\text{IV}}(\gamma) = 0$, can be seen as choosing, under all causal models compatible with the distribution, the model with the least amount confounding – when using $E(Y - \gamma^\top X)^2 - E(Y - \gamma^\top_{\text{OLS}} X)^2$ as a measure for confounding.

Equation (2.1) is quadratic in α, the resulting constraint is, in general, non-convex. In Section 2.3.5, we develop a computationally efficient procedure that provably solves the optimization problem at low computational cost. Other choices of tests are possible, too, but may result in even harder optimization problems.

In Section 2.3.1, we briefly introduce the setup and assumptions. In Section 2.3.2, we specify a class of asymptotically consistent tests for $\mathcal{H}_0(\alpha)$. In Section 2.3.3 we formally define the PULSE estimator. In Section 2.3.4, we show that the PULSE estimator is well-defined by proving that it is equivalent to a solvable convex quadratically constrained quadratic program which we denote by the primal PULSE. In Section 2.3.5, we utilize duality theory and derive an alternative representation which we denote by the dual PULSE. This representation yields a computationally feasible algorithm and shows that the PULSE estimator is a K-class estimator with a data-driven κ. Proofs of results in this section can be found in Appendix A.5 unless stated otherwise.

2.3.1. Setup and Assumptions

In the following sections we again let $(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{A})$ consist of $n \geq \min\{d, q\}$ row-wise independent and identically distributed copies of (Y, X, H, A) generated in accordance with the SEM in Equation (2.1). The structural equation of interest is $Y = \gamma_0^\intercal X + \eta_0^\intercal H + \beta_0^\intercal A + \varepsilon_Y$. Assume that we have some non-sample information about which $d_2 = d - d_1$ and $q_2 = q - q_1$ coefficients of γ_0 and β_0, respectively, are zero. As in Section 2.2, we let the subscript $*$ denote the variables and coefficients that are non-zero according to the non-sample information but to simplify notation, we drop the $*$ subscript from Z, $\mathbf{Z}$ and α_0; that is, we write $Z = [X_*^\intercal \; A_*^\intercal]^\intercal \in \mathbb{R}^{d_1+q_1}$, $\mathbf{Z} = [\mathbf{X}_* \; \mathbf{A}_*] \in \mathbb{R}^{n\times(d_1+q_1)}$ and $\alpha_0 := (\gamma_{0,*}^\intercal, \beta_{0,*}^\intercal)^\intercal :\in \mathbb{R}^{d_1+q_1}$. That is, $Y = \alpha_0^\intercal Z + U_Y$, where $U_Y = \alpha_{0,-*}^\intercal Z_{-*} + \eta_0^\intercal H + \varepsilon_Y$. If the non-sample information is true, then $U_Y = \eta_0^\intercal H + \varepsilon_Y$.

We define a setup as being under- just- and over-identified by the degree of over-identification $q_2 - d_1$ being negative, equal to zero and positive, respectively. That is, the number of excluded exogenous variables A_{-*} being less, equal or larger than the number of included endogenous variables X_* in the target equation.

We assume that the global assumptions of Assumption 2.1 from Section 2.2.1 still hold. Furthermore, we will make use of the following situational assumptions

Assumption 2.4. *(a)* $A \perp\!\!\!\perp U_Y$*;* *(b)* $E[A] = 0$*.*

Assumption 2.5. ε *has non-degenerate marginals.*

Assumption 2.6. *(a)* $\mathbf{Z}^\intercal\mathbf{Z}$ *is of full rank;* *(b)* $\mathbf{A}^\intercal\mathbf{Z}$ *is of full rank.*

Assumption 2.7. $[\mathbf{Z}\ \mathbf{Y}]$ *is of full column rank.*

Assumption 2.8. $E[AZ^\intercal]$ *is of full rank.*

Assumption 2.4.(a) holds if our non-sample information is true, and the instrument set A is independent of all unobserved endogenous variables H_i which directly

affect the target Y. This holds, for example, if the latent variables are source nodes, that is, they have no parents in the causal graph of the corresponding SEM. Assumption 2.4.(b) can be achieved by centering the data. Strictly speaking, this introduces a weak dependence structure in the observations, which is commonly ignored. Alternatively, one can perform sample splitting. For more details on this assumption and the possibility of relaxing it, see Remark 2.1. Assumption 2.6.(a) ensures that K-class estimators for $\kappa < 1$ are well-defined, regardless of the over-identification degree. In the under-identified setup, Assumption 2.6.(b) yields that there exists a subspace of solutions minimizing $l^n_{\mathrm{IV}}(\alpha)$. In the just- and over-identified setup this assumption ensures that $l^n_{\mathrm{IV}}(\alpha)$ has a unique minimizer given by the two-stage least squares estimator $\hat{\alpha}^n_{\mathrm{TSLS}} := (\mathbf{Z}^\intercal P_{\mathbf{A}} \mathbf{Z})^{-1} \mathbf{Z}^\intercal P_{\mathbf{A}} \mathbf{Y}$. Assumption 2.7 is used to ensure that the ordinary least square objective function $l^n_{\mathrm{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z})$ is strictly positive, such that division by this function is always well-defined. Assumptions 2.5 and 2.8 ensure that various limiting arguments are valid. In the just- and over-identified setup Assumption 2.8 is known as the rank condition for identification.

2.3.2. Testing for Vanishing Correlation

We now introduce a class of tests for the null hypothesis $\mathcal{H}_0(\alpha) : \mathrm{Corr}(A, Y - Z\alpha) = 0$ that have point-wise asymptotic level and pointwise asymptotic power. These tests will allow us to define the corresponding PULSE estimator. When Assumption 2.7 holds we can define $T^c_n : \mathbb{R}^{d_1+q_1} \to \mathbb{R}$ by

$$T^c_n(\alpha) := c(n) \frac{l^n_{\mathrm{IV}}(\alpha)}{l^n_{\mathrm{OLS}}(\alpha)} = c(n) \frac{\|P_{\mathbf{A}}(\mathbf{Y} - \mathbf{Z}\alpha)\|_2^2}{\|\mathbf{Y} - \mathbf{Z}\alpha\|_2^2},$$

where $c(n)$ is a function that will typically scale linearly in n. Let us denote the $1-p$ quantile of the central Chi-Squared distribution with q degrees of freedom by $Q_{\chi^2_q}(1-p)$. By standard limiting theory we can test $\mathcal{H}_0(\alpha)$ in the following manner.

Lemma 2.1 (Level and power of the test)**.** *Let Assumptions 2.4, 2.5 and 2.7 hold and assume that $c(n) \sim n$ as $n \to \infty$. For any $p \in (0,1)$ and any fixed α, the statistical test rejecting the null hypothesis $\mathcal{H}_0(\alpha)$ if $T^c_n(\alpha) > Q_{\chi^2_q}(1-p)$, has point-wise asymptotic level p and point-wise asymptotic power of 1 against all alternatives as $n \to \infty$.*

Remark 2.1. Assumption 2.4.(b), $E[A] = 0$, is important for the test statistic to be asymptotic pivotal under the null hypothesis, that is, to ensure that the asymptotic distribution of $T^c_n(\alpha)$ does not depend on the model parameters except for q. We can drop this assumption if we change the null hypothesis to $\mathcal{H}_0(\alpha) : E[A(Y - Z^\intercal\alpha)] = 0$ and add the assumption that $E[U_Y] = 0$. Furthermore, if we are in the just- or over-identified setup and Assumption 2.8 holds, both of these hypotheses are under their respective assumptions equivalent to $\tilde{\mathcal{H}}_0(\alpha) : \alpha = \alpha_0$.

That is, the test in Lemma 2.1 becomes an asymptotically consistent test for the causal coefficient. ◦

Depending on the choice of $c(n)$, this class contains several tests, some of which are well known. With $c(n) = n - q + Q_{\chi^2_q}(1 - p_{\min})$, for example, one recovers a test that is equivalent to the asymptotic version of the Anderson-Rubin test (Anderson and Rubin, 1950). We make this connection precise in Remark A.2 in Appendix A.7. The Anderson-Rubin test is robust to weak instruments in the sense that the limiting distribution of the test-statistic under the null-hypothesis is not affected by weak instrument asymptotics; see, e.g. Staiger and Stock (1997) and Stock et al. (2002).[6] For weak instruments, the confidence region may be unbounded with large probability; see Dufour (1997). Moreira (2009) show that the test suffers from loss of power in the over-identified setting.

To simplify notation, we will from now on work with the choice $c(n) = n$ and define the acceptance region with level $p_{\min} \in (0, 1)$ as $\mathcal{A}_n(1 - p_{\min}) := \{\alpha \in \mathbb{R}^{d_1+q_1} : T_n(\alpha) \leq Q_{\chi^2_q}(1 - p_{\min})\}$, where $T_n(\alpha)$ corresponds to the choice $c(n) = n$.

2.3.3. The PULSE Estimator

For any level $p_{\min} \in (0, 1)$, we formally define the PULSE estimator of Equation (2.1) by letting the feasible set be given by the acceptance region $\mathcal{A}_n(1-p_{\min})$ of $\mathcal{H}_0(\alpha)$ using the test of Lemma 2.1. That is, we consider

$$\hat{\alpha}^n_{\text{PULSE}}(p_{\min}) := \begin{array}{ll} \arg\min_\alpha & l^n_{\text{OLS}}(\alpha) \\ \text{subject to} & T_n(\alpha) \leq Q_{\chi^2_q}(1 - p_{\min}). \end{array} \tag{2.2}$$

In general, this is a non-convex optimization problem (Boyd and Vandenberghe, 2004) as the constraint function is non-convex, see the blue contours in Figure 2.1(left). From Figure 2.1(right) we see that in the given example the problem nevertheless has a unique and well-defined solution: the smallest level-set of l^n_{OLS} with a non-empty intersection of the acceptance region $\{\alpha : T_n(\alpha) \leq Q_{\chi^2_q}(1-p_{\min})\}$ intersects with the latter region in a unique point. In Section 2.3.4, we prove that this is not a coincidence: Equation (2.2) has a unique solution that coincides with the solution of a strictly convex, quadratically constrained quadratic program (QCQP) with a data-dependent constraint bound. In Section 2.3.5, we further derive an equivalent Lagrangian dual problem. This has two important implications. (1) It allows us to construct a computationally efficient procedure to compute a solution of the non-convex problem above, and (2), it shows that the PULSE estimator can be written as K-class estimators.

Estimators with similar constraints albeit different optimization objective have been studied by Gautier et al. (2018). In Remark A.3 in Appendix A.7 we briefly discuss the connection to pre-test estimators. Furthermore, any method

[6] Weak instrument asymptotics is a model scheme where the instrument strength tends to zero at a rate of $n^{-1/2}$, i.e., the reduced form structural equation for the endogenous variables is given by $\mathbf{X} = \mathbf{A}n^{-1/2}\Pi_X + \varepsilon\Gamma_X^{-1}$.

for inverting the test, see, e.g., Davidson and MacKinnon (2014), yields a valid confidence set including the proposed point estimator (given that the method outputs the point estimator when the acceptance region is empty).

2.3.4. Primal Representation of PULSE

We now derive a QCQP representation of the PULSE problem, which we call the primal PULSE. For all $t \geq 0$ define the empirical primal minimization problem (Primal.t.n) by

$$\begin{array}{ll} \text{minimize}_\alpha & l^n_{\text{OLS}}(\alpha; \mathbf{Y}, \mathbf{Z}) \\ \text{subject to} & l^n_{\text{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A}) \leq t. \end{array} \tag{2.3}$$

We drop the dependence of $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{A}$ and refer to the objective and constraint functions as $l^n_{\text{OLS}}(\alpha)$ and $l^n_{\text{IV}}(\alpha)$. The following lemma shows that under suitable assumptions these problems are solvable, strictly convex QCQP problems satisfying Slater's condition.

Lemma 2.2 (Unique solvability of the primal)**.** *Let Assumption 2.6 hold. It holds that $\alpha \mapsto l^n_{\text{OLS}}(\alpha)$ and $\alpha \mapsto l^n_{\text{IV}}(\alpha)$ are strictly convex and convex, respectively. Furthermore, for any $t > \inf_\alpha l^n_{\text{IV}}(\alpha)$ it holds that the constrained minimization problem (Primal.t.n) has a unique solution and satisfies Slater's condition. In the under- and just-identified setup the constraint bound requirement is equivalent to $t > 0$ and in the over-identified setup to $t > l^n_{\text{IV}}(\hat{\alpha}^n_{\text{TSLS}})$, where $\hat{\alpha}^n_{\text{TSLS}} = (\mathbf{Z}^\intercal P_{\mathbf{A}} \mathbf{Z})^{-1} \mathbf{Z}^\intercal P_{\mathbf{A}} \mathbf{Y}$.*

We restrict the constraint bounds to $D_{\text{Pr}} := (\inf_\alpha l^n_{\text{IV}}(\alpha), l^n_{\text{IV}}(\hat{\alpha}^n_{\text{OLS}})]$. Considering t that are larger than $\inf_\alpha l^n_{\text{IV}}(\alpha)$ ensures that the problem (Primal.t.n) is uniquely solvable and furthermore that Slater's condition is satisfied (see Lemma 2.2 above). Slater's condition will play a role in Section 2.3.5 when establishing a sufficiently strong connection with its corresponding dual problem for which we can derive a (semi-)closed form solution. Constraint bounds greater than or equal to $l^n_{\text{IV}}(\hat{\alpha}^n_{\text{OLS}})$ yield identical solutions. Whenever well-defined, let $\hat{\alpha}^n_{\text{Pr}} : D_{\text{Pr}} \to \mathbb{R}^{d_1+q_1}$ denote the constrained minimization estimator given by the solution to the (Primal.t.n) problem

$$\hat{\alpha}^n_{\text{Pr}}(t) := \begin{array}{ll} \arg\min_\alpha & l^n_{\text{OLS}}(\alpha) \\ \text{subject to} & l^n_{\text{IV}}(\alpha) \leq t. \end{array} \tag{2.4}$$

We now prove that for a specific choice of t, the PULSE and the primal PULSE yield the same solutions. Define $t^\star_n(p_{\min})$ as the data-dependent constraint bound given by

$$t^\star_n(p_{\min}) := \sup\{t \in (\inf_\alpha l^n_{\text{IV}}(\alpha), l^n_{\text{IV}}(\hat{\alpha}^n_{\text{OLS}})] : T_n(\hat{\alpha}^n_{\text{Pr}}(t)) \leq Q_{\chi^2_q}(1 - p_{\min})\}. \tag{2.5}$$

If $t^\star_n(p_{\min}) > -\infty$ or equivalently $t^\star_n(p_{\min}) \in D_{\text{Pr}}$ we define the primal PULSE problem and its solution by (Primal.$t^\star_n(p_{\min})$.n) and $\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p_{\min}))$. The following theorem yields conditions for when the solutions to the primal PULSE and PULSE problems coincide.

Theorem 2.2 (Primal representation of PULSE)**.** *Let $p_{\min} \in (0,1)$ and Assumptions 2.6 and 2.7 hold and assume that $t_n^\star(p_{\min}) > -\infty$. If it holds that $T_n(\hat{\alpha}_{\Pr}^n(t_n^\star(p_{\min}))) \leq Q_{\chi_q^2}(1-p_{\min})$, then the PULSE problem has a unique solution given by the primal PULSE solution. That is, $\hat{\alpha}_{\text{PULSE}}^n(p_{\min}) = \hat{\alpha}_{\Pr}^n(t_n^\star(p_{\min}))$.*

In the proof of Theorem 2.3, we show that $t_n^\star(p_{\min}) > -\infty$ is a sufficient to guarantee that $T_n(\hat{\alpha}_{\Pr}^n(t_n^\star(p_{\min}))) \leq Q_{\chi_q^2}(1-p_{\min})$. The sufficiency of $t_n^\star(p_{\min}) > -\infty$ is postponed to the latter proof as it easily follows from the dual representation. Hence, we have shown that finding the PULSE estimator, i.e., finding a solution to the non-convex PULSE problem, is equivalent to solving the convex QCQP primal PULSE for a data dependent choice of $t_n^\star(p_{\min})$.[7] However, $t_n^\star(p_{\min})$ is still unknown. Figure 2.1 shows an example of the equivalence in Theorem 2.2. Figure 2.1(right) shows that the level set of $l_{\text{IV}}(\alpha) = t^\star(p_{\min})$ intersects the optimal level curve of $l_{\text{OLS}}^n(\alpha)$ in the same point given by minimizing over the constraint $T_n(\alpha) \leq Q_{\chi_q^2}(1-p_{\min})$.

The set of solutions to the primal problem $\{\hat{\alpha}_{\Pr}^n(t) : t \in D_{\Pr}\}$ can in the just- and over-identified setup be visualized as an (in general) non-linear path in $\mathbb{R}^{d_1+q_1}$ between the TSLS estimator ($t = l_{\text{IV}}^n(\hat{\alpha}_{\text{TSLS}}^n)$) and the OLS estimator ($t = l_{\text{IV}}^n(\hat{\alpha}_{\text{OLS}}^n)$) (see also Rothenhäusler et al., 2021). Theorem 2.2 yields that the PULSE estimator ($t = t_n^\star(p_{\min})$) then seeks the estimator 'closest' to the OLS estimator along this path that does not yield a rejected test of simultaneous vanishing correlation between the resulting prediction residuals and the exogenous variables A, see Figure 2.1. The path of possible solutions is not necessarily a straight line (see black line); thus, in general, the PULSE estimator is different from the affine combination of OLS and TSLS estimators studied by e.g. Judge and Mittelhammer (2012).

In the under-identified setup, the TSLS end point corresponding to $t = \min_\alpha l_{\text{IV}}^n(\alpha)$ is instead given by the point in the IV solution space $\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\text{IV}}^n(\alpha) = 0\}$ with the smallest mean squared prediction residuals.

2.3.5. Dual Representation of PULSE

In this section, we derive a dual representation of the primal PULSE problem which we will denote the dual PULSE problem. This specific dual representation allows for the construction of a binary search algorithm for the PULSE estimator and yields that PULSE is a member of the K-class estimators with stochastic κ-parameter.

For any penalty parameter $\lambda \geq 0$ we define the dual problem (Dual.λ.n) by

$$\text{minimize} \quad l_{\text{OLS}}^n(\alpha) + \lambda l_{\text{IV}}^n(\alpha). \tag{2.6}$$

Whenever Assumption 2.6.(a) holds, i.e., $\mathbf{Z}^\intercal\mathbf{Z}$ is of full rank, then for any $\lambda \geq 0$ the solution to (Dual.λ.n) coincides with the K-class estimator with $\kappa = \lambda/(1+\lambda) \in$

[7] Given that value, we can use a numerical QCQP solver to calculate the PULSE estimate.

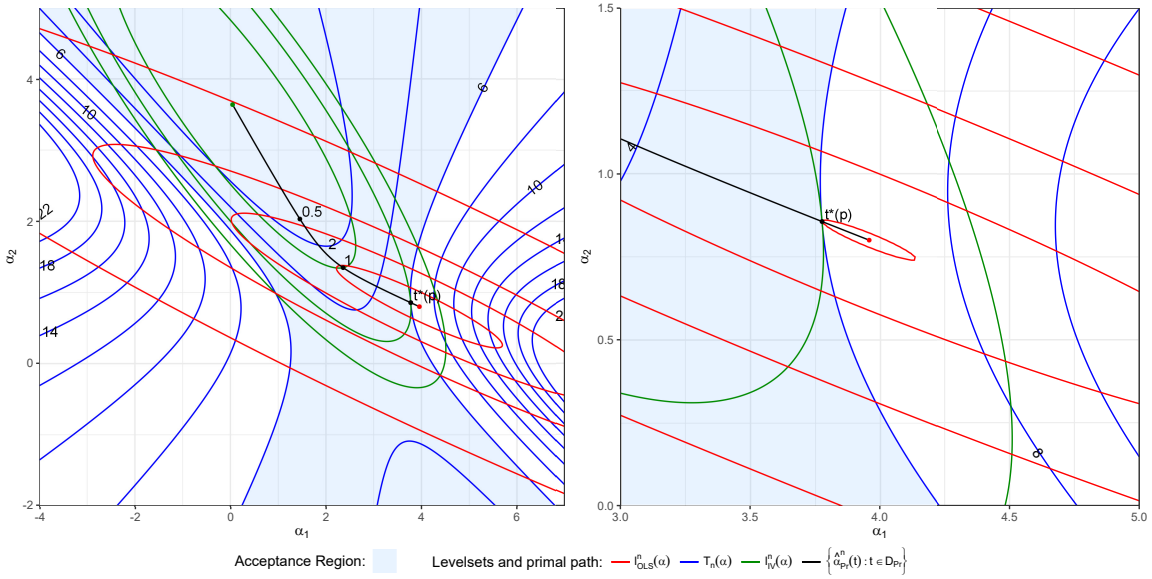


Figure 2.1: Illustrations of the level sets of l^n_{OLS} (red contours), the proposed test-statistic T_n (blue contours) and l^n_{IV} (green contours) in a just-identified setup. The example is generated with a two dimensional anchor $A = (A_1, A_2)$, one of which is included, and one included endogenous variable X, i.e., $Y = \alpha_1 X + \alpha_2 A_1 + H + \varepsilon_Y$ with $(\alpha_1, \alpha_2) = (1, 1)$. Both illustrations show level sets from the same setup, but they use different scales. The black text denotes the level of the test-statistic contours. In this setup, the PULSE constraint bound, the rejection threshold of the test with $p_{\min} = 0.05$, is $Q_{\chi^2_2}(0.95) \approx 5.99$. The blue level sets of T_n are non-convex. The sublevel set of the test, corresponding to the acceptance region, is illustrated by the blue area. In the right plot, we see that the smallest level set of l^n_{OLS} that has a non-empty intersection with the $Q_{\chi^2_q}(1 - p_{\min})$-sublevel set of T_n is a singleton (black dot, $t^*(p)$). This shows that in this example the PULSE problem is solvable and has a unique solution. The l^n_{IV} level set that intersects this singleton is exactly the $t^\star_n(p_{\min})$-level set of l^n_{IV}, illustrating the statement of Theorem 2.2 in that the primal PULSE with that choice of t solves the PULSE problem. The black line visualizes the solutions $\{\hat{\alpha}^n_{\text{Pr}}(t) : t \in D_{\text{Pr}}\}$. The black points and corresponding text labels indicates which constraint bound t yields the specific point. In general, the class of primal solutions does not coincide with the class of convex combinations of the OLS and the TSLS estimators.

$[0, 1)$, see Proposition 2.1. That is,

$$\hat{\alpha}^n_{\text{K}}(\kappa) = (\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Y}$$

solves (Dual.λ.n). Henceforth, let $\hat{\alpha}^n_{\text{K}}(\lambda)$ denote the solution to (Dual.λ.n), i.e., in a slight abuse of notation we will denote the solution to (Dual.λ.n) by $\hat{\alpha}^n_{\text{K}}(\lambda)$,

such that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{K}}(\kappa)$ for $\kappa = \lambda/(1+\lambda)$. We refer to these two representations as the K-class estimator with penalty parameter λ and parameter κ, respectively. The usage of κ or λ as argument should clarify which notation we refer to.

Under Assumption 2.6.(b) we have that the minimum of $l^n_{\mathrm{IV}}(\alpha)$ is attainable (see the proof of Lemma 2.2). Hence, let the solution space for the minimization problem $\min_\alpha l^n_{\mathrm{IV}}(\alpha)$ be given by

$$\mathcal{M}_{\mathrm{IV}} := \arg\min_\alpha l^n_{\mathrm{IV}}(\alpha) = \{\alpha \in \mathbb{R}^{d_1+q_1} : l^n_{\mathrm{IV}}(\alpha) = \min_{\alpha'} l^n_{\mathrm{IV}}(\alpha')\}. \tag{2.7}$$

In the under-identified setup ($q_2 < d_1$), $\mathcal{M}_{\mathrm{IV}}$ is a $(d_1 - q_2)$-dimensional subspace of $\mathbb{R}^{d_1+q_1}$ and in the just- and over-identified setup it holds that $\mathcal{M}_{\mathrm{IV}} = \{\hat{\alpha}^n_{\mathrm{TSLS}}\}$.

We now prove that, in the generic case, K-class estimators for $\lambda \in [0, \infty)$ are different from the TSLS estimator. This result may not come as a surprise, but we include it as we need the result later and have not found it elsewhere.

Lemma 2.3 (K-class estimators and TSLS differ)**.** *Assume that we are in the just- or over-identified setup and $n > q$. Furthermore, assume that ε has density with respect to Lebesgue measure and that the coefficient matrix B of the SEM in Equation* (2.1) *is lower triangular. If the rank conditions of Assumption 2.6 hold almost surely, then it almost surely holds, that all K-class estimators with penalty parameter $\lambda \in [0, \infty)$ differ from the TSLS estimator, i.e., $\hat{\alpha}^n_{\mathrm{TSLS}} \notin \{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\}$.*

We conjecture that the corresponding statement holds in the under-identified setup and without the lower triangular assumption on B, too. That is, $\mathcal{M}_{\mathrm{IV}} \cap \{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\} = \emptyset$ holds almost surely. We therefore introduce this as an assumption.

Assumption 2.9. *No K-class estimator $\hat{\alpha}^n_{\mathrm{K}}(\kappa)$ with $\kappa \in [0, 1)$, is a member of $\mathcal{M}_{IV}$.*

Furthermore, when imposing that Assumption 2.9 holds we also have that the K-class estimators differ from each other.

Corollary 2.1 (K-class estimators differ)**.** *Let Assumptions 2.6 and 2.9 hold. If $\lambda_1, \lambda_2 \geq 0$ with $\lambda_1 \neq \lambda_2$, then $\hat{\alpha}^n_{\mathrm{K}}(\lambda_1) \neq \hat{\alpha}^n_{\mathrm{K}}(\lambda_2)$.*

The above corollary is proven as Corollary A.1 in Appendix A.4. We now show that the class of K-class estimators with penalty parameter $\lambda \geq 0$, i.e., $\kappa \in [0, 1)$, coincides with the class of constrained minimization-estimators that minimize the primal problems with constraint bounds $t > \min_\alpha l^n_{\mathrm{IV}}(\alpha)$.

Lemma 2.4 (Connecting the primal and dual)**.** *If Assumptions 2.6, 2.7 and 2.9 hold, then both of the following statements hold. (a) For any $t \in D_{\mathrm{Pr}}$, there exists a unique $\lambda(t) \geq 0$ such that (Primal.t.n) and (Dual.$\lambda(t)$.n) have the same unique solution. (b) For any $\lambda \geq 0$, there exists a unique $t(\lambda) \in D_{\mathrm{Pr}}$ such that (Primal.$t(\lambda)$.n) and (Dual.λ.n) have the same unique solution.*

Lemma 2.4 tells us that, under appropriate assumptions, $\{\hat{\alpha}^n_{\mathrm{K}}(\kappa) : \kappa \in [0,1)\} = \{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\} = \{\hat{\alpha}^n_{\mathrm{Pr}}(t) : t \in D_{\mathrm{Pr}}\}$. In words, we have recast the K-class estimators with $\kappa \in [0,1)$ as the class of solutions to the primal problems previously introduced. That the minimizers of $l^n_{\mathrm{IV}}(\alpha)$ are different from all the K-class estimators with penalty $\lambda \geq 0$ (or $\kappa \in [0,1)$) guarantees that when representing a K-class problem in terms of a constrained optimization problem it satisfies Slater's condition.

We are now able to show the main result of this section. The PULSE estimator $\hat{\alpha}^n_{\mathrm{PULSE}}(p_{\min})$ solves a K-class problem (Dual.λ.n) and can therefore be seen as a K-class estimator with a data-dependent parameter. To see this, let us define the dual PULSE penalty parameter, i.e., the dual analogue of the primal PULSE constraint $t^\star_n(p_{\min})$ as

$$\lambda^\star_n(p_{\min}) := \inf\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q_{\chi^2_q}(1 - p_{\min})\}. \tag{2.8}$$

If $\lambda^\star_n(p_{\min}) < \infty$, we define the dual PULSE problem by (Dual.$\lambda^\star_n(p_{\min})$.$n$) with solution $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p)) = \arg\min_{\alpha \in \mathbb{R}^{d_1+q_1}} l^n_{\mathrm{OLS}}(\alpha) + \lambda^\star_n(p_{\min}) l^n_{\mathrm{IV}}(\alpha)$.

Theorem 2.3 (Dual representation of PULSE)**.** *Let $p_{\min} \in (0,1)$ and Assumptions 2.6, 2.7 and 2.9 hold. If $\lambda^\star_n(p_{\min}) < \infty$, then it holds that $t^\star_n(p_{\min}) > -\infty$ and $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) = \hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min})) = \hat{\alpha}^n_{\mathrm{PULSE}}(p_{\min})$.*

Thus, the PULSE estimator seeks to minimize the K-class penalty λ, i.e., to pull the estimator along the K-class path $\{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\}$ as close to the ordinary least square estimator as possible. Furthermore, the statement implies that the PULSE estimator is a K-class estimator with data-driven penalty $\lambda^\star_n(p_{\min})$ or, equivalently, parameter $\kappa = \lambda^\star_n(p_{\min})/(1 + \lambda^\star_n(p_{\min}))$. Given a finite dual PULSE penalty parameter $\lambda^\star_n(p_{\min})$ we can, by utilizing the closed form solution of the K-class problem, represent the PULSE estimator in the following form:

$$\begin{aligned}\hat{\alpha}^n_{\mathrm{PULSE}}(p_{\min}) &= \hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) \\ &= (\mathbf{Z}^\intercal(\mathbf{I} + \lambda^\star_n(p_{\min})P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{I} + \lambda^\star_n(p_{\min})P_{\mathbf{A}})\mathbf{Y}.\end{aligned}$$

However, to the best of our knowledge, $\lambda^\star_n(p_{\min})$ has no known closed form, so the above expression cannot be computed in closed-form either. In Section 2.3.5.1, we prove that the PULSE penalty parameter $\lambda^\star_n(p_{\min})$ can be approximated with arbitrary precision by a simple binary search procedure.

The following lemma provides a necessary and sufficient (in practice checkable) condition for when the PULSE penalty parameter $\lambda^\star_n(p_{\min})$ is finite.

Lemma 2.5 (Infeasibility of the dual representation)**.** *Let $p_{\min} \in (0,1)$ and Assumptions 2.6, 2.7 and 2.9 hold. In the under- and just-identified setup we have that $\lambda^\star_n(p_{\min}) < \infty$. In the over-identified setup it holds that $\lambda^\star_n(p_{\min}) < \infty \iff T_n(\hat{\alpha}^n_{\mathrm{TSLS}}) < Q_{\chi^2_q}(1 - p_{\min})$. This is not guaranteed to hold as the event that $\mathcal{A}_n(1 - p_{\min}) = \emptyset$ can have positive probability.*

Thus, under suitable regularity assumptions Lemma 2.5 yields that our dual representation of the PULSE estimator always holds in the under- and just-identified setup. It furthermore yields a sufficient and necessary condition for the dual representation to be valid in the over-identified setup, namely that the TSLS is in the interior of the acceptance region. Furthermore, this condition is possibly violated in the over-identified setup with non-negligible probability.

2.3.5.1. Binary Search for the Dual Parameter

The key insight allowing for a binary search procedure for $\lambda_n^\star(p_{\min})$ is that the mapping $\lambda \mapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ is monotonically decreasing.

Lemma 2.6 (Monotonicity of the losses and test statistic)**.** *When Assumption 2.6.(a) holds the maps* $[0,\infty) \ni \lambda \mapsto l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *and* $[0,\infty) \ni \lambda \mapsto l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *are monotonically increasing and monotonically decreasing, respectively. Consequently, if Assumption 2.7 holds, we have that the map* $[0,\infty) \ni \lambda \mapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *is monotonically decreasing. Furthermore, if Assumption 2.9 also holds, these monotonicity statements can be strengthened to strictly decreasing and strictly increasing.*

The above lemma is proven as Lemma A.1 in Appendix A.4. If the OLS solution is not strictly feasible in the PULSE problem, then $\lambda_n^\star(p_{\min})$ indeed is the smallest penalty parameter for which the test-statistic reaches a p-value of exactly $p_{\min}$; see Lemma A.2 in Appendix A.4.

We propose the binary search algorithm presented in Algorithm A.1 in Appendix A.2, that can approximate a finite $\lambda_n^\star(p_{\min})$ with arbitrary precision. We terminate the binary search (see line 2) if $\lambda_n^\star(p_{\min})$ is not finite, in which case we have no computable representation of the PULSE estimator. It is possible to improve this algorithm in the under- and just-identified setup, by initializing $\ell_{\max}$ as the quantity given by Equation (A.28) in the proof of Lemma 2.5. This initialization removes the need for the first while loop in (lines 4–6). We now prove that Algorithm A.1 achieves the required precision and is asymptotically correct.

Lemma 2.7. *Let* $p_{\min} \in (0,1)$ *and Assumptions 2.6 and 2.7 hold. If it holds that* $\lambda_n^\star(p_{\min}) < \infty$*, then* $\lambda_n^\star(p_{\min})$ *can be approximated with arbitrary precision by the binary search Algorithm A.1, that is,* $\mathrm{Binary.Search}(N, p_{\min}) - \lambda_n^\star(p_{\min}) \to 0$*, as* $N \to \infty$*.*

2.3.6. Algorithm and Consistency

The dual representation of the PULSE estimator is not guaranteed to be well-defined in the over-identified setup. In particular, it is not well-defined if the TSLS is outside the interior of the acceptance region (which corresponds to a p-value of less than or equal to $p_{\min}$). In this case, we propose to output a warning. This can be helpful information for the user since it may indicate a model misspecification.

For example, if the true relationship is in fact nonlinear, and one considers an over-identified case (e.g., by constructing different transformations of the instrument), even the TSLS may be rejected when erroneously considering a linear model; see Keane (2010) and Mogstad and Wiswall (2010). For any $p_{\min} \in (0,1)$ we can still define an always well-defined modified PULSE estimator $\hat{\alpha}^n_{\text{PULSE+}}(p_{\min})$ as $\hat{\alpha}^n_{\text{PULSE}}(p_{\min})$ if the dual representation is feasible and some other consistent estimator $\hat{\alpha}^n_{\text{ALT}}$ (such as the TSLS, LIML or Fuller estimator) otherwise. That is, we define

$$\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) := \begin{cases} \hat{\alpha}^n_{\text{PULSE}}(p_{\min}), & \text{if } T_n(\hat{\alpha}^n_{\text{TSLS}}) < Q_{\chi^2_q}(1-p_{\min}) \\ \hat{\alpha}^n_{\text{ALT}}, & \text{otherwise.} \end{cases}$$

Similarly to the case of an empty rejection region, we also output a warning for the case when the OLS estimator is accepted. This may, but does not have to, indicate weak instruments. Thus, we have the algorithm presented as Algorithm A.2 in Appendix A.2 for computing the PULSE+ estimator.

We now prove that the PULSE+ estimator consistently estimates the causal parameter in the just- and over-identified setting. Assume that we choose a consistent estimator $\hat{\alpha}^n_{\text{ALT}}$ (under standard regularity assumptions, this is satisfied for the TSLS).[8] We can then show that, under mild conditions, the PULSE+ estimator, too, is a consistent estimator of α_0.

Theorem 2.4 (Consistency of PULSE+)**.** *Consider the just- or over-identified setup and let $p_{\min} \in (0,1)$. If Assumptions 2.4 and 2.6 to 2.9 hold almost surely for all $n \in \mathbb{N}$ and $\hat{\alpha}^n_{\text{ALT}}$ consistently estimates α_0, then $\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) \xrightarrow{P} \alpha_0$, when $n \to \infty$.*

We believe that a similar statement also holds in the under-identified setting, see Section A.8.3.

2.4. Simulation Experiments

In Appendix A.8 we conduct an extensive simulation study investigating the finite sample behaviour of the PULSE estimator. The concept of weak instruments is central to our analysis. An introduction to weak instruments can be found in Appendix A.10. Here we give a brief overview of the study and the observations.

2.4.1. Distributional Robustness

The theoretical results on distributional robustness proved in Section 2.2 translate to finite data. The experiments of Section A.8.1 shows that even for small sample sizes, K-class estimators outperform both OLS and TSLS for a certain range of interventions, matching the theoretical predictions with increasing sample size. In Section A.8.3, we furthermore consider an under-identified setting.

[8] Since $\hat{\alpha}^n_{\text{TSLS}} = \alpha_0 + (n^{-1}\mathbf{Z}^\intercal\mathbf{A}(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1}n^{-1}\mathbf{A}^\intercal\mathbf{Z})^{-1}n^{-1}\mathbf{Z}^\intercal\mathbf{A}(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1}n^{-1}\mathbf{A}^\intercal\mathbf{U}_Y$.

2.4.2. Estimating Causal Effects

When focusing on the estimation of a causal effect in an identified setting, our simulations show that there are several settings where PULSE outperforms the Fuller and TSLS estimators in terms of mean squared error (MSE). In univariate simulation experiments, such settings are characterized by weakness of instruments and weak confounding (endogeneity). The characterization becomes more involved in multivariate settings, but is similar in that PULSE outperforms all other methods for small confounding strengths, an effect amplified by the weakness of instruments. Below we detail the univariate simulation setup and refer the reader to Appendix A.8 for further details and the multivariate simulation experiments mentioned above.

2.4.2.1. Univariate Model.

We first compared performance measures of the estimators in a univariate instrumental variable model. As seen in Hahn and Hausman (2002) and Hahn et al. (2004), we consider structural equation models of the form

$$A := A \in \mathbb{R}^q, \quad X := A^\intercal \bar{\xi} + U_X \in \mathbb{R}, \quad Y := X\gamma + U_Y \in \mathbb{R},$$

where $A \sim \mathcal{N}(0, I)$ and $A \perp\!\!\!\perp (U_X, U_Y)$ with $\left(\begin{smallmatrix} U_X \\ U_Y \end{smallmatrix}\right) \sim \mathcal{N}\left(\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right)$. Furthermore, we let $\gamma = 1$ and $\bar{\xi}^\intercal = (\xi, ..., \xi) \in \mathbb{R}^q$, where $\xi > 0$ is chosen according to the theoretical R^2-coefficient. We consider the following simulation scheme: for each $q \in \{1, 2, 3, 4, 5, 10, 20, 30\}$, $\rho \in \{0.1, 0.2, ..., 0.9\}$, $R^2 \in \{0.0001, 0.001, 0.01, 0.1, 0.3\}$ and $n \in \{50, 100, 150\}$, we simulate n-samples from the above system and calculate the OLS, TSLS, Fuller(1), Fuller(4) and PULSE ($p_{\min} = 0.05$) estimates; see Section A.8.2.1.

Figure 2.2 contains illustrations of the relative change in square-root mean squared error (RMSE) estimated from 15000 repetitions. On the horizontal axis we have plotted the average first stage F-test as a measure of weakness of instruments; see Appendix A.10 for further details. A test for $H_0 : \bar{\xi} = 0$, i.e., for the relevancy of instruments, at a significance level of 5%, has different rejection thresholds in the range $[1.55, 4.04]$ depending on n and q. The vertical dashed line corresponds to the smallest rejection threshold of 1.55 and the dotted line corresponds to the 'rule of thump' threshold of 10. Note that the lowest possible negative relative change is -1 and a positive relative change means that PULSE is better.

In Appendix A.11, further illustrations of e.g. the relative change in mean bias and variance of the estimators are presented. We also conducted the simulations for setups with combinations of $\gamma \in \{-1, 0\}$, components of $\bar{\xi}$ chosen negatively, with random flipped sign in each coordinate and for negative ρ (not shown but available in the folder 'Plots' in the code repository). The results with respect to MSE are similar to those shown in Figure 2.2, while the bias comparison changes depending on the setup.

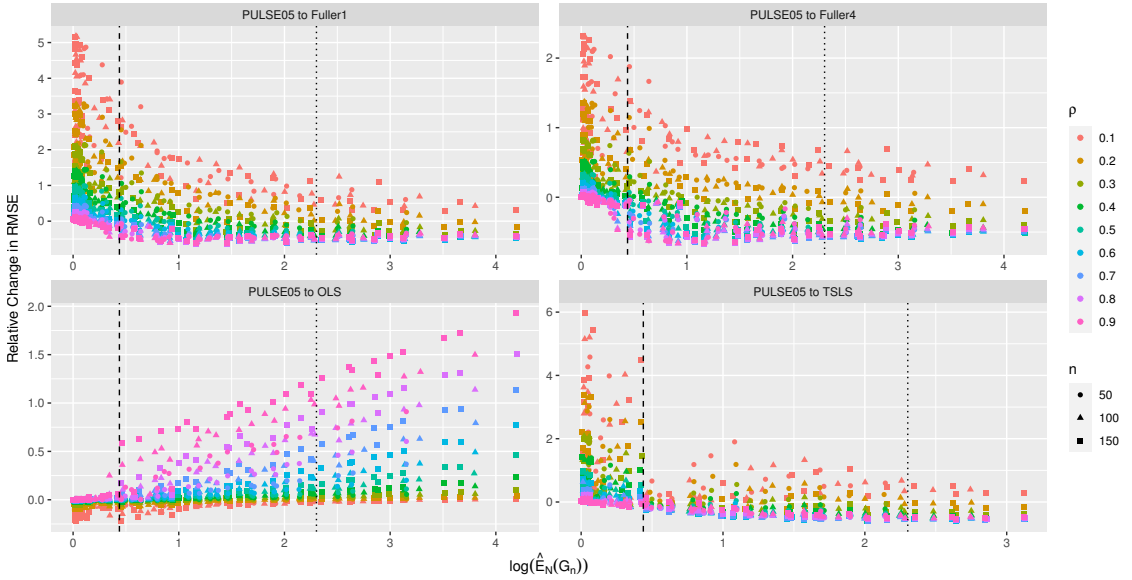


Figure 2.2: Illustrations of the relative change in RMSE.

We observe that there are settings, in which the PULSE is superior to TSLS, Fuller(1) and Fuller(4) in terms of MSE. This is particularly often the case in weak instrument settings ($\hat{E}_N(G_n) < 10$) for low confounding strength ($\rho \leq 0.2$). Furthermore, as we tend towards the weakest instrument setting considered, we also see a gradual shift in favour of PULSE for higher confounding strengths. In these settings with weak instruments and low confounding we also see that OLS is superior to the PULSE in terms of MSE. However, for large confounding setups PULSE is superior to OLS in terms of both bias and MSE and this superiority increases as the instrument strength increases. The PULSE is generally more biased than the Fuller and TSLS estimators but less biased than OLS. However, in the settings with weak instruments and low confounding the bias of PULSE and OLS is comparable. In summary, the PULSE is in these settings more biased but its variance is so small that it is MSE superior to the Fuller and TSLS estimators.

2.5. Empirical Applications

We now consider three classical instrumental variable applications (see Albouy (2012) and Buckles and Hungerman (2013) for discussions on the underlying assumptions).

(i) "Does compulsory school attendance affect schooling and earnings?" by Angrist and Krueger (1991). This paper investigates the effects of education on wages. The endogenous effect of education on wages are remedied by instrumenting education on quarter of birth indicators.

(ii) "Using geographic variation in college proximity to estimate the return to schooling" by Card (1993). This paper also investigates the effects of

education on wages. In this paper education is instrumented by proximity to college indicator.

(iii) "The colonial origins of comparative development: An empirical investigation" by Acemoglu et al. (2001). This paper investigates the effects of extractive institutions (proxied by protection against expropriation) on the gross domestic product (GDP) per capita. The endogeneity of the explanatory variables are remedied by instrumenting protection against expropriation on early European settler mortality rates.

We have applied the different estimators OLS, TSLS, PULSE, and Fuller to the classical data sets Acemoglu et al. (2001), Angrist and Krueger (1991) and Card (1993). All models considered in Angrist and Krueger (1991) and Card (1993), where we estimate the effect on years of education on wages, using quarter of birth and proximity to colleges as instruments, respectively, the OLS estimates are not rejected by our test statistic and PULSE outputs the OLS estimates; see Appendix A.9 for futher details. This may be either due to weak endogeneity (weak confounding), or that the test has insufficient power to reject the OLS estimates due to either weak instruments or severe over-identification.

2.5.1. Acemoglu et al. (2001)

The dataset of Acemoglu et al. (2001) consists of 64 observations, each corresponding to a different country for which mortality rate estimates encountered by the first European settlers are available. The endogenous target of interest is log GDP per capita (in 1995). The main endogenous regressor in the dataset is an index of expropriation protection (averaged over 1985–1995), i.e., protection against expropriation of private investment by the respective governments. The average expropriation protection is instrumented by the settler mortality rates. We consider eight models M1–M8 which correspond to the models presented in column (1)–(8) in Table 4 of Acemoglu et al. (2001). Model M1 is given by the reduced form structural equations

$$\log \mathrm{GDP} = \mathrm{avexpr} \cdot \gamma + \mu_1 + U_1, \quad \mathrm{avexpr} = \log \mathrm{em4} \cdot \delta + \mu_2 + U_2,$$

where avexpr is the average expropriation protection, em4 is the settler mortality rates, μ_1 and μ_2 are intercepts and U_1 and U_2 are possibly correlated, unobserved noise variables. In model M2 we additionally introduce an included exogenous regressor describing the country latitude. In model M3 and M4 we fit model M1 and M2, respectively, on a dataset where we have removed Neo-European countries, Australia, Canada, New Zealand and the United States. In model M5 and M6 we fit model M1 and M2, respectively, on a dataset where we have removed observations from the continent of Africa. In model M7 and M8 we again fit model M1 and M2, respectively, but now also include three exogenous indicators for the continents Africa, Asia and other.

Table 2.1 shows the OLS and TSLS estimates (which replicate the values from the study), as well as the Fuller(4) and PULSE estimates for the linear effect of the average expropriation protection on log GDP. In model M1, for example, we see that the PULSE estimate suggests that the average expropriation risk linear effect on log GDP is 0.6583 which is 26% larger than the OLS estimate but 34% smaller than TSLS estimate. In models M5–M8, the OLS estimates are not rejected by the Anderson-Rubin test, so the PULSE estimates coincide with the OLS estimates.

We can also use this example to illustrate the robustness property of K-class estimators; see Theorem 2.1. Even though interventional data are not available, we can consider the mean squared prediction error when holding out the observations with the most extreme values of the instrument. Depending on the degree of generalization, we indeed see that the PULSE and Fuller tend to outperform OLS or TSLS in terms of mean squared prediction error on the held out data; see Section A.9.3 for further details.

Table 2.1: The estimated return of expropriation protection on log GDP per capita.

Model	OLS	TSLS	FUL	PULSE	Message	Test	Threshold
M1	0.5221	0.9443	0.8584	0.6583	–	5.991	5.991
M2	0.4679	0.9957	0.8457	0.5834	–	7.815	7.815
M3	0.4868	1.2812	0.9925	0.7429	–	5.991	5.991
M4	0.4709	1.2118	0.9268	0.6292	–	7.815	7.815
M5	0.4824	0.5780	0.5573	0.4824	OLS Acc.	1.180	5.991
M6	0.4658	0.5757	0.5476	0.4658	OLS Acc.	1.155	7.815
M7	0.4238	0.9822	0.7409	0.4238	OLS Acc.	10.772	11.071
M8	0.4013	1.1071	0.7059	0.4013	OLS Acc.	9.755	12.592

Note: Point estimates for the return of expropriation protection on log GDP per capita. The OLS and TSLS values coincide with the ones shown in Acemoglu et al. (2001). The right columns show the values of the test statistic (evaluated in the PULSE estimates) and the test rejection thresholds. The '–' indicates that OLS is not accepted and TSLS is not rejected.

2.6. Summary and Future Work

We have proved that a distributional robustness property similar to the one shown for anchor regression (Rothenhäusler et al., 2021) fully extends to general K-class estimators of possibly non-identifiable structural parameters in a general linear structural equation model that allows for latent endogenous variables. We have further proposed a novel estimator for structural parameters in linear structural equation models. This estimator, called PULSE, is derived as the solution to a minimization problem, where we seek to minimize mean squared prediction error constrained to a confidence region for the causal parameter. Even though this region is non-convex, we have shown that the corresponding optimization

problem allows for a computationally efficient algorithm that approximates the above parameter with arbitrary precision using a simple binary search procedure. In the under-identified setting, this estimator extends existing work in the machine learning literature that considers invariant subsets or the best predictive sets among them: PULSE is applicable even in situations when no invariant subsets exist. We have proved that this estimator can also be written as a K-class estimator with data-driven κ-parameter, which lies between zero and one. Simulation experiments show that in various settings with weak instruments and weak confounding, PULSE outperforms other estimators such as the Fuller(4) estimator. We thus regard PULSE as an interesting alternative for estimating causal effects in instrumental variable settings. It is easy to interpret and automatically provides the user feedback in case that the OLS is accepted (which may be an indication that the instruments are too weak) or that the TSLS is outside the acceptance region (which may indicate a model misspecification). We have applied the different estimators to classical data sets and have seen that, indeed, K-class estimators tend to be more distributionally robust than OLS or TSLS.

There are several further directions that we consider worthwhile investigating. This includes better understanding of finite sample properties and for the identified setups, the study of loss functions other than MSE. It would be helpful, in particular with respect to real world applications, to understand to which extent similar principles can be applied to models allowing for a time structure of the error terms. We believe that the simple primal form of PULSE could make it applicable for model classes that are more complex than linear models (see also Christiansen et al., 2021). Our procedure can be combined with other tests and it could furthermore be interesting to find efficient optimization procedures for tests that are robust with respect to weak instruments, such as Kleibergen's K-statistic (Kleibergen, 2002), for example. In an under-identified setting, the causal parameters are not identified but the solutions obtained by optimizing predictability under invariance might be promising candidates for models that generalize well to distributional shifts.

Acknowledgements

We are grateful to Trine Boomsma, Peter Bühlmann, Rune Christiansen, Steffen Lauritzen, Nicolai Meinshausen, Whitney Newey, Cosma Shalizi, and Nikolaj Thams for helpful discussions. We thank the editor and two anonymous referees for helpful and constructive comments. MEJ and JP were supported by the Carlsberg Foundation; JP was, in addition, supported by a research grant (18968) from VILLUM FONDEN.

Chapter 3

A Causal Framework for Distribution Generalization

Joint work with

Rune Christiansen, Niklas Pfister, Nicola Gnecco,
and Jonas Peters

Abstract

We consider the problem of predicting a response Y from a set of covariates X when test and training distributions differ. Since such differences may have causal explanations, we consider test distributions that emerge from interventions in a structural causal model, and focus on minimizing the worst-case risk. Causal regression models, which regress the response on its direct causes, remain unchanged under arbitrary interventions on the covariates, but they are not always optimal in the above sense. For example, for linear models and bounded interventions, alternative solutions have been shown to be minimax prediction optimal. We introduce the formal framework of distribution generalization that allows us to analyze the above problem in partially observed nonlinear models for both direct interventions on X and interventions that occur indirectly via exogenous variables A. It takes into account that, in practice, minimax solutions need to be identified from data. Our framework allows us to characterize under which class of interventions the causal function is minimax optimal. We prove sufficient conditions for distribution generalization and present corresponding impossibility results. We propose a practical method, NILE, that achieves distribution generalization in a nonlinear IV setting with linear extrapolation. We prove consistency and present empirical results.

Keywords: Distribution generalization, causality, worst-case risk, distributional robustness, invariance, domain adaptation

3.1. Introduction

Large-scale learning systems, particularly those focusing on prediction tasks, have been successfully applied in various domains of application. Since inference is usually done during training time, any difference between training and test

distribution poses a challenge for prediction methods (Arjovsky et al., 2019; Csurka, 2017; Pan and Yang, 2010; Quionero-Candela et al., 2009). Dealing with these differences is of great importance in several fields such as environmental sciences, where methods need to extrapolate both in space and time. Tackling this problem requires restrictions on how the distributions may differ, since, clearly, generalization becomes impossible if the test distribution may be arbitrary. Given a response Y and some covariates X, several existing procedures aim to find a minimax function f which minimizes the worst-case risk $\sup_{P \in \mathcal{N}} \mathbb{E}_P[(Y - f(X))^2]$ across distributions contained in a small neighborhood $\mathcal{N}$ of the training distribution. The neighborhood $\mathcal{N}$ should be representative of the difference between the training and test distributions, and often mathematical tractability is taken into account, too (Abadeh et al., 2015; Sinha et al., 2018). A typical approach is to define a ρ-ball of distributions $\mathcal{N}_\rho(P_0) := \{P : D(P, P_0) \leq \rho\}$ around the (empirical) training distribution P_0, with respect to some divergence measure D, such as the Kullback-Leibler divergence (Bagnell, 2005; Hu and Hong, 2013). While some divergence functions only consider distributions with the same support as P_0, the Wasserstein distance allows for a neighborhood of distributions around P_0 with possibly different supports (Abadeh et al., 2015; Blanchet et al., 2019; Esfahani and Kuhn, 2018; Sinha et al., 2018).

In our analysis, we do not start from a divergence measure, but instead model the difference between training and test distribution using the concept of interventions (Pearl, 2009; Peters et al., 2017). We believe that for many problems this provides a useful description of distributional changes. We will see that, depending on the considered setup, this approach allows to find models that perform well even on test distributions which would be considered far away from the training distribution in any commonly used metric. For this class of distributions, causal regression models appear naturally because of the following well-known observation. A prediction model, which uses only the direct causes of the response Y as covariates, is invariant under interventions on variables other than Y: the conditional distribution of Y given its causes does not change (this principle is known, e.g., as invariance, autonomy or modularity) (Aldrich, 1989; Haavelmo, 1944; Pearl, 2009). Such a causal regression model yields the minimal worst-case risk when considering all interventions on variables other than Y (e.g., Rojas-Carulla et al., 2018a, Theorem 1, Appendix). It has therefore been suggested to use causal models in problems of distributional shifts (Arjovsky et al., 2019; Heinze-Deml and Meinshausen, 2021; Magliacane et al., 2018; Meinshausen, 2018; Pfister et al., 2021; Rojas-Carulla et al., 2018a; Schölkopf et al., 2012). In practice, however, not all relevant causal variables might be observed. One may further argue that causal methods are too conservative in that the interventions which induce the test distributions may not be arbitrarily strong. Instead, methods which focus on a trade-off between predictability and causality have been proposed for linear models (Pfister et al., 2019; Rothenhäusler et al., 2021), see also Section 3.5.1. Anchor regression (Rothenhäusler et al., 2021) is shown to be predictive optimal under a set of bounded interventions.

In this work, we introduce the general framework of distribution generalization,

which permits a unifying perspective on the potentials and limitations of applying causal concepts to the problem of generalizing regression models from training to test distribution. In particular, we use it to characterize the relationship between a minimax optimal solution and the causal function, and to classify settings under which the minimax solution is identifiable from the training distribution.

3.1.1. Further Related Work

The field of distributional robustness or out-of-distribution generalization aims to develop procedures that are robust to changes between training and test distribution. This problem has been actively studied from an empirical perspective in machine learning research, for example, in image classification by using adversarial attacks, where small digital (Goodfellow et al., 2014) or physical (Evtimov et al., 2017) perturbations of pictures can deteriorate the performance of a model. Arguably, these procedures are not yet fully understood theoretically. A more theoretical perspective is given by the previously mentioned minimization of a worst-case risk across distributions contained in a neighborhood of the training distribution, in our case, distributions generated by interventions.

Our framework includes the problems of multi-task learning, domain generalization and transfer learning (Baxter, 2000; Caruana, 1997; Mansour et al., 2009; Quionero-Candela et al., 2009) (see Section 3.2.4 for more details), with a focus on minimizing the worst-case risk. In settings of covariate shift (e.g., Shimodaira, 2000; Sugiyama and Müller, 2005; Sugiyama et al., 2008), one usually assumes that the training and test distribution of the covariates are different, while the conditional distribution of the response given the covariates remains invariant (Ben-David et al., 2010; Bickel et al., 2009; Daume III and Marcu, 2006; Muandet et al., 2013). Sometimes, it is additionally assumed that the support of the training distribution covers that of the test distribution (Shimodaira, 2000). In this work, the conditional distribution of the response given the covariates is allowed to change between interventions, due to the existence of hidden confounders, and we consider settings where the test observations lie outside the training support.

Data augmentation methods have become successful techniques, e.g. in image classification, to adapt prediction procedures to such types of distribution shifts. These methods increase the diversity of the training data by changing the geometry and the color of the images (e.g., by rotation, cropping or changing saturation) (Shorten and Khoshgoftaar, 2019; Zhang et al., 2018). This allows the user to create models that generalize better to unseen environments (e.g., Volpi et al., 2018). We view these approaches as ways to enlarge the support of the covariates, which, as our results show, comes with theoretical advantages, see Section 3.4.

Minimizing the worst-case risk is considered in robust methods (El Ghaoui et al., 2003; Kim et al., 2006), too. It can also be formulated in terms of minimizing the regret in a multi-armed bandit problem (Auer et al., 2002; Bartlett et al., 2008; Lai and Robbins, 1985). In that setting, the agent can choose the distribution which generates the data. In our setting, though, we do not assume to have control over

the interventions, and, hence, neither over the distribution of the sampled data.

3.1.2. Contribution and Structure

This work contains four main contributions: (1) A novel framework for analyzing the problem of generalization from training to test distribution, using the notion of distribution generalization (Section 3.2). (2) Results elucidating the relationship between a causal function and a minimax solution (Section 3.3). (3) Sufficient conditions which ensure distribution generalization, along with corresponding impossibility results (Section 3.4). (4) A practical method, called NILE ('Nonlinear Intervention-robust Linear Extrapolator'), which learns a minimax solution from i.i.d. observational data (Section 3.5).

Our framework describes how structural causal models can be used as technical devices for modeling plausible test distributions. It further allows us to formally define distribution generalization, which describes the ability to identify generalizing regression models (i.e., minimax solutions) from the observational distribution. While it is well known that the causal function is minimax optimal under the set of all interventions on the covariates (e.g., Rojas-Carulla et al., 2018a), we extend this result in several ways, for example, by allowing for hidden variables and by characterizing more general sets of interventions under which the causal function is minimax optimal. We further derive conditions on the model class, the observational distribution and the family of interventions under which distribution generalization is possible, and present impossibility results proving the necessity of some of these conditions. For example, we show that strong assumptions on the functional relationship between X and Y are needed whenever the interventions extend the training support of X. An example of such an assumption is to consider the class of differentiable functions that linearly extrapolate outside the support of X. For that model class, we propose the explicit method NILE, which obtains distribution generalization by exploiting a nonlinear instrumental variables setup. We show that our method learns a minimax solution which corresponds to the causal function. We prove consistency and compare our algorithm to state-of-the art approaches empirically.

We believe that our results shed some light on the potential merits of using causal concepts in the context of generalization. The framework allows us to make first steps towards answering when it can be beneficial to use non-causal functions for prediction under interventions, and what might happen under misspecification of the intervention class. Our results also formalize in which sense methods that generalize in the linear case – such as IV and anchor regression (Rothenhäusler et al., 2021) – can be extended to nonlinear settings. Further, our framework implies impossibility statements for multi-task learning that relate to existing results (Ben-David et al., 2010).

Our code is available as an `R`-package at `https://runesen.github.io/NILE`; scripts generating all our figures and results can be found at the same url. Additional supporting material is given in the online appendix. Appendix B.1 shows

how to represent several causal models in our framework. Appendix B.2 summarizes existing results on identifiability in IV models. Appendix B.3 provides details on the test statistic that we use for NILE. Appendix B.4 contains additional experiments. All proofs are provided in Appendix B.5.

3.2. Framework

For a real-valued response $Y \in \mathbb{R}$ and predictors $X \in \mathbb{R}^d$, we consider the problem of identifying a regression function that works well not only on the training data, but also under perturbed distributions that we will model by interventions.

3.2.1. Modeling Intervention-induced Distributions

We require a model that is able to model an observational distribution of (X, Y) (as training distribution) and the distribution of (X, Y) under a class of interventions on (parts of) X (as test distribution). We will do so by means of a structural causal model (SCM) (Bollen, 1989; Pearl, 2009). More precisely, denoting by $H \in \mathbb{R}^q$ some additional (unobserved) variables, we consider the SCM

$$H := \varepsilon_H, \;\; X := h_2(H, \varepsilon_X), \;\; Y := f(X) + h_1(H, \varepsilon_Y), \tag{3.1}$$

where the assignments for H, X and Y consist of q, d and 1 coordinate(s), respectively. Here, f, h_1 and h_2 are measurable functions, and the innovation terms ε_X, ε_Y and ε_H are independent vectors with possibly dependent coordinates. Two comments are in order. First, the joint distribution of (X, Y) is constrained only by requiring that X and $h_1(H, \varepsilon_Y)$ enter the assignment for Y additively. This constraint affects the allowed conditional distributions of Y given X, but does not make any restriction on the marginal distributions of either X or Y. Second, we only use the above SCM as a technical device for modeling training and test distributions, by considering interventions on X or A (introduced in Section 3.2.3), for which we are analyzing the predictive performance of different models – similarly to how one could have considered a ball around the training distribution. We therefore only require the SCM to correctly (a) model the training-distribution, and (b) induce the test-distributions through interventions. Any other causal implications of the SCM, such as causal orderings between variables, causal effects or counterfactual statements, are not assumed to be correctly specified. As such, our framework includes a wide range of cases, including situations where training and test distribution come from interventions in an SCM with a different structure than (3.1), where, for example, some of the variables in X are not ancestors but descendants of Y. To see whether our framework applies, one needs to check if the considered training and test distributions can be equivalently expressed as interventions in a model of our form. If the structure of the true data generating SCM is known, this can be done by directly transforming the SCM and the interventions. The following remark shows an example of such a transformation

and may be interesting to readers with a special interest in causality. It can be skipped at first reading.

Remark 3.1 (Transforming causal models)**.** Assume that the training distribution is induced by the following SCM

$$X_1 \coloneqq \varepsilon_1, \quad X_2 \coloneqq k(Y) + \varepsilon_2, \quad Y \coloneqq f(X_1) + \varepsilon_3,$$

with $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q$, and that we consider test distributions arising from shift interventions on X_2. This set of training and test distributions can be equivalently modeled by the reduced SCM

$$H \coloneqq \varepsilon_3, \quad X \coloneqq h_2(H, (\varepsilon_1, \varepsilon_2)), \quad Y \coloneqq f(X_1) + H,$$

with $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q$, and where h_2 is defined by

$$h_2(H, (\varepsilon_1, \varepsilon_2)) \coloneqq (\varepsilon_1, k(f(\varepsilon_1) + H) + \varepsilon_2).$$

Both SCMs induce the same observational distribution over (X_1, X_2, Y) and shift interventions on X_2 in the original SCM correspond to shift interventions on $X = (X_1, X_2)$ in the reduced SCM (where only the second coordinate is shifted). Our framework can then be used, for example, to give sufficient conditions under which generalization (formally defined below) is possible, see Proposition 3.7 and 3.8.

It is not always possible to transform an SCM into our reduced form, and it might also happen that the transformed interventions are not covered by our framework. For example, we do not allow for direct interventions on Y in the original model. In other cases, where the original SCM may contain additional hidden variables, even interventions on (parts of) X in the original SCM may translate into interventions on H in the reduced SCM, and are therefore not covered. Details and a more general treatment are provided in Appendix B.1. ∘

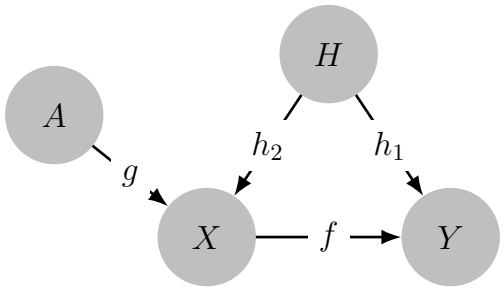
Sometimes, the vector of covariates X contains variables, which are independent of H, that enter into the assignments of the other covariates additively and cannot be used for the prediction (e.g., because they are not observed during testing). If such covariates exist, it can be useful to explicitly distinguish them from the remaining predictors. We will denote them by A and call them exogenous variables. Such variables are interesting for several reasons. (i) We will see that in general, interventions on A lead to intervention distributions with desirable properties for distribution generalization, see Section 3.4.4. (ii) Some of our results rely on the function f being identifiable from the observational distribution, see Assumption 3.1 below. The variables A can be used to state explicit conditions for identifiability. Under additional assumptions, for example, they can be used as instrumental variables (e.g., Bowden and Turkington, 1985; Greene, 2003), a well-established tool for recovering f from the observational distribution of (X, Y, A). (iii) The variable A can be used to model a covariate that is not observed under testing. It can also be used to index tasks (which we discuss at the end of Section 3.2.4). In

the remainder of this work, we therefore consider a slightly larger class of SCMs that also includes exogenous variables A. It contains the SCM (3.1) as a special case.[1] We derive results for settings with and without exogenous variables A.

3.2.2. Model

Formally, we consider a response $Y \in \mathbb{R}^1$, covariates $X \in \mathbb{R}^d$, exogenous variables $A \in \mathbb{R}^r$, and unobserved variables $H \in \mathbb{R}^q$. Let further $\mathcal{F} \subseteq \{f : \mathbb{R}^d \to \mathbb{R}\}$, $\mathcal{G} \subseteq \{g : \mathbb{R}^r \to \mathbb{R}^d\}$, $\mathcal{H}_1 \subseteq \{h_1 : \mathbb{R}^{q+1} \to \mathbb{R}\}$ and $\mathcal{H}_2 \subseteq \{h_2 : \mathbb{R}^{q+d} \to \mathbb{R}^d\}$ be fixed sets of measurable functions. Moreover, let $\mathcal{Q}$ be a collection of probability distributions on $\mathbb{R}^{d+1+r+q}$, such that for all $Q \in \mathcal{Q}$ it holds that if $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$, then $\varepsilon_X, \varepsilon_Y, \varepsilon_A$ and ε_H are jointly independent, and for all $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ it holds that $\xi_Y := h_1(\varepsilon_H, \varepsilon_Y)$ and $\xi_X := h_2(\varepsilon_H, \varepsilon_X)$ have mean zero.[2] Let $\mathcal{M} := \mathcal{F} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ denote the model class. Every model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ then specifies an SCM by[3]

$$
\begin{aligned}
A &:= \varepsilon_A \\
H &:= \varepsilon_H \\
X &:= g(A) + h_2(H, \varepsilon_X) \\
Y &:= f(X) + h_1(H, \varepsilon_Y)
\end{aligned}
$$



with $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$, where the assignments for A, H , X and Y consist of r, q, d and 1 coordinate(s), respectively. For each model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, we refer to f as the *causal function* (for the pair (X, Y)), and denote by $\mathbb{P}_M$ the joint distribution over the observed variables (X, Y, A). We assume that this distribution has finite second moments. If no exogenous variables A exist, one can think of the function g as being constant. A model M that correctly models the training and test distributions will be referred to as the 'true model'.

3.2.3. Interventions

Each SCM $M \in \mathcal{M}$ can now be modified by the concept of interventions (e.g., Pearl, 2009; Peters et al., 2017). An intervention corresponds to replacing one or more of the structural assignments of the SCM (see Section 3.4.2 for details on the types of interventions considered in this paper). For example, we intervene on some of the covariates X by replacing the corresponding assignments with, e.g., a Gaussian random vector that is independent of the other noise variables. Importantly, an

[1] This follows from choosing A as an independent noise variable and a constant g.

[2] This can be assumed w.l.o.g. if $\mathcal{F}$ and $\mathcal{G}$ are closed under addition and scalar multiplication, and contain the constant function.

[3] For an appropriate choice of h_2, the model includes settings in which (parts of) A directly influence Y.

intervention on some of the variables does not change the assignment of any other variable. In particular, an intervention on X does not change the conditional distribution of Y, given X and H (this is an instance of the invariance property mentioned in Section 3.1) but it may change the conditional distribution of Y, given X.

The problems addressed in this work require us to simultaneously consider several different SCMs that are all subject to the same (set of) interventions. Formally, we therefore regard an intervention i as a mapping from the model class $\mathcal{M}$ into a (possibly larger) set of SCMs, which takes as input a model $M \in \mathcal{M}$ and outputs another model $M(i)$ over variables (X^i, A^i, Y^i, H^i), the intervened model. We do not need to assume that the intervened model $M(i)$ belongs to the model class $\mathcal{M}$, but we require that $M(i)$ induces a joint distribution over (X^i, Y^i, A^i, H^i)[4] with finite second moments. We denote the corresponding distribution over the observed (X^i, Y^i, A^i) by $\mathbb{P}_{M(i)}$, and use $\mathcal{I}$ for a collection of interventions. In our work, the test distributions are modeled as distributions generated by these types of intervened models, and the set $\mathcal{I}$ therefore indexes the set of test distributions. We will be interested in the mean squared prediction error on each test distribution i, formally written as $\mathbb{E}_{M(i)}[(Y - f(X))^2]$. (In this work, we consider a univariate Y, but writing $\mathbb{E}[\|Y - f_\diamond(X)\|^2_{\mathbb{R}^d}] = \sum_{j=1}^d \mathbb{E}[(Y_j - f_{\diamond,j}(X))^2]$, most our results extend straight-forwardly to a d-dimensional response.)

The support of random variables under interventions will play an important role for the analysis of distribution generalization. Throughout this paper, $\text{supp}^M(Z)$ denotes the support of the random variable $Z \in \{A, X, H, Y\}$ under the distribution induced by the SCM $M \in \mathcal{M}$. Moreover, $\text{supp}^M_{\mathcal{I}}(Z)$ denotes the union of $\text{supp}^{M(i)}(Z)$ over all interventions $i \in \mathcal{I}$. We call a collection of interventions on Z *support-reducing* (w.r.t. M) if $\text{supp}^M_{\mathcal{I}}(Z) \subseteq \text{supp}^M(Z)$ and *support-extending* (w.r.t. M) if $\text{supp}^M_{\mathcal{I}}(Z) \not\subseteq \text{supp}^M(Z)$. Whenever it is clear from the context which model is considered, we may drop the indication of M altogether and simply write $\text{supp}(Z)$.

3.2.4. Distribution Generalization

Let $\mathcal{M}$ be a fixed model class, let $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ and let $\mathcal{I}$ be a class of interventions. In this work, we aim to find a function $f^* : \mathbb{R}^d \to \mathbb{R}$, such that the predictive model $\hat{Y} = f^*(X)$ has low worst-case risk over all test distributions induced by the interventions $\mathcal{I}$ in model M. We therefore consider, for the true M, the optimization problem

$$\arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right], \tag{3.2}$$

where $\mathbb{E}_{M(i)}$ is the expectation in the intervened model $M(i)$. In general, this optimization problem is neither guaranteed to have a solution, nor is the solution,

[4] If the context does not allow for any ambiguity, we omit the superscript i.

if it exists, ensured to be unique. Whenever a solution f^* to (3.2) exists, we refer to it as a *minimax solution* (for model M w.r.t. $(\mathcal{F}, \mathcal{I})$).

Depending on the model class $\mathcal{M}$, there may be several models $\tilde{M} \in \mathcal{M}$ that induce the observational distribution $\mathbb{P}_M$, that is, the same distribution over the observed variables A, X and Y, but do not agree with M on all intervention distributions induced by $\mathcal{I}$. Thus, each such model induces a potentially different minimax problem with different solutions. Given knowledge only of $\mathbb{P}_M$, it is therefore generally not possible to identify a solution to (3.2). In this paper, we study conditions on $\mathcal{M}$, $\mathbb{P}_M$ and $\mathcal{I}$, under which this becomes possible. More precisely, we aim to characterize under which conditions $(\mathbb{P}_M, \mathcal{M})$ admits distribution generalization to $\mathcal{I}$.

Definition 3.1 (Distribution generalization)**.** $(\mathbb{P}_M, \mathcal{M})$ *is said to* admit distribution generalization to $\mathcal{I}$*, or simply to* admit generalization to $\mathcal{I}$*, if for every* $\varepsilon > 0$ *there exists a function* $f^*_\varepsilon \in \mathcal{F}$ *such that, for all models* $\tilde{M} \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$*, it holds that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f^*_\varepsilon(X))^2 \right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f_\diamond(X))^2 \right] \right| \leq \varepsilon. \tag{3.3}$$

Distribution generalization does not require the existence of a minimax solution in $\mathcal{F}$ (which would require further assumptions on the function class $\mathcal{F}$) and instead focuses on whether an approximate solution can be identified based only on the observational distribution $\mathbb{P}_M$. If, however, there exists a function $f^* \in \mathcal{F}$ which, for every $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, is a minimax solution for $\tilde{M}$ w.r.t. $(\mathcal{F}, \mathcal{I})$, then, in particular, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$.

Our framework also includes several settings of multitask learning (MTL) and domain adaptation (Quionero-Candela et al., 2009), where one often assumes to observe different training tasks. In MTL, one is then interested in using the different tasks to improve the predictive performance on either one or all training tasks – this is often referred to as asymmetric and symmetric MTL, respectively. In our framework, such a setup can be modeled using a categorical variable X. If, however, one is interested in predicting on an unseen task or if one does not know which of the observed tasks the new test data come from, one may instead use a categorical A with support-extending or support-reducing interventions, respectively.

3.3. Minimax Solutions and the Causal Function

To address the question of distribution generalization, we first study properties of the minimax optimization problem (3.2). In the simplest case, where $\mathcal{I}$ consists only of the trivial intervention, that is, $\mathbb{P}_M = \mathbb{P}_{M(i)}$, we are looking for the best predictor on the observational distribution. In that case, the minimax solution is attained at any conditional mean function, $f^* : x \mapsto \mathbb{E}[Y|X = x]$ (provided that $f^* \in \mathcal{F}$). For larger classes of interventions, however, the conditional mean

may become sub-optimal in terms of prediction. To see this, it is instructive to decompose the risk under an intervention. Since the structural assignment for Y remains unchanged for all interventions that we consider in this work, it holds for all $f_\diamond \in \mathcal{F}$ and all interventions i on either A or X that

$$\begin{aligned}\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = &\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \\ &+ 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))].\end{aligned}$$

Here, the middle term does not depend on i since $\xi_Y = h_1(H, \varepsilon_Y)$ remains fixed. We call the intervention i

confounding-removing	if for all models $M \in \mathcal{M}$ it holds that $X \perp\!\!\!\perp H$, under $M(i)$.

For such an intervention, we have that $\xi_Y \perp\!\!\!\perp X$ under $\mathbb{P}_{M(i)}$, and hence, since $\mathbb{E}_M[\xi_Y] = 0$, the last term in the above equation vanishes. Therefore, if $\mathcal{I}$ consists only of confounding-removing interventions, the causal function is a solution to the minimax problem (3.2). The following proposition shows that an even stronger statement holds: The causal function is already a minimax solution if $\mathcal{I}$ contains at least one confounding-removing intervention on X.

Proposition 3.1 (Confounding-removing interventions on X)**.** *Let $\mathcal{I}$ be a set of interventions on X or A such that there exists at least one $i \in \mathcal{I}$ that is confounding-removing. Then, the minimal worst-case risk is attained at a confounding-removing intervention, and the causal function f is a minimax solution.*

We now prove that, in a linear setting, the causal function is also minimax optimal if the interventions create unbounded variability in all directions of the covariance matrix of X.

Proposition 3.2 (Unbounded interventions on X with linear $\mathcal{F}$)**.** *Let $\mathcal{F}$ be the class of all linear functions, and let $\mathcal{I}$ be a set of interventions on X or A s.t. $\sup_{i\in\mathcal{I}} \lambda_{\min}\big(\mathbb{E}_{M(i)}\big[XX^\top\big]\big) = \infty$, where $\lambda_{\min}$ denotes the smallest eigenvalue. Then, the causal function f is the unique minimax solution.*

The unbounded eigenvalue condition above is satisfied if $\mathcal{I}$ is the set of all shift interventions on X. These interventions, formally defined in Section 3.4.2.2, appear in linear IV models and recently gained further attention in the causal community (Rothenhäusler et al., 2021; Sani et al., 2020). The proposition above considers a linear function class $\mathcal{F}$; in this way, shift interventions are related to linear models.

Even if the causal function f does not solve the minimax problem (3.2), the difference between the minimax solution and the causal function cannot be arbitrarily large. The following proposition shows that the worst-case L_2-distance between f and any function $f_\diamond$ that performs better than f (in terms of worst-case risk) can be bounded by a term which is related to the strength of the confounding.

Proposition 3.3 (Difference between causal function and minimax solution)**.** *Let $\mathcal{I}$ be a set of interventions on X or A. Then, for any function $f_\diamond \in \mathcal{F}$ which satisfies that*

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

it holds that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4\,\mathrm{Var}_M[\xi_Y].$$

Even though the difference can be bounded, it may be non-zero, and one may benefit from choosing a function that differs from the causal function f. This choice, however, comes at a cost: it relies on the fact that we know the class of interventions $\mathcal{I}$. In general, being a minimax solution is not entirely robust with respect to misspecification of $\mathcal{I}$. In particular, if the set $\mathcal{I}_2$ of interventions describing the test distributions is misspecified by a set $\mathcal{I}_1 \neq \mathcal{I}_2$, then the considered minimax solution with respect to $\mathcal{I}_1$ may perform worse than the causal function on the test distributions.

Proposition 3.4 (Properties of the minimax solution under mis-specified interventions)**.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be any two sets of interventions on X, and let $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$, it holds that*

$$\sup_{i\in\mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big] \leq \sup_{i\in\mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f(X))^2\Big].$$

If $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$, however, it can happen (even if $\mathcal{F}$ is linear) that

$$\sup_{i\in\mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big] > \sup_{i\in\mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f(X))^2\Big].$$

The second part of the proposition should be understood as a non-robustness property of non-causal minimax solutions. Improvements on the causal function are possible in situations, where one has reasons to believe that the test distributions do not stem from a set of interventions that is much larger than the specified set.

3.4. Distribution Generalization

As described in Section 3.2.4, we consider a fixed model class $\mathcal{M}$ containing the true (but unknown) model M, and let $\mathcal{I}$ be a class of interventions. By definition, the optimizer of the minimax problem (3.2) depends on the true model M. Section 3.3 relates this optimizer to the causal function f, whose knowledge, too, requires knowing M. In practice, however, we do not have access to the true model M, but only to its observational distribution $\mathbb{P}_M$. This motivates the notion of distribution generalization, see (3.3). In words, it states that approximate minimax solutions (which depend on the intervention distributions $\mathbb{P}_{M(i)}$, $i \in \mathcal{I}$) are identified from the observational distribution $\mathbb{P}_M$. This holds true, in particular, if the intervention distributions themselves are identified from $\mathbb{P}_M$.

Intervention on	$\text{supp}_{\mathcal{I}}(X)$	Assumptions	Result
X (well-behaved)	$\subseteq \text{supp}(X)$	Ass. 3.1	Proposition 3.7
X (well-behaved)	$\not\subseteq \text{supp}(X)$	Ass. 3.1 and 3.2	Proposition 3.8
A	$\subseteq \text{supp}(X)$	Ass. 3.1 and 3.3	Proposition 3.12
A	$\not\subseteq \text{supp}(X)$	Ass. 3.1, 3.2 and 3.3	Proposition 3.12

Table 3.1: Summary of conditions under which generalization is possible. Corresponding impossibility results are shown in Propositions 3.6, 3.11 and 3.13.

Proposition 3.5 (Sufficient conditions for distribution generalization)**.** *Assume that for all $\tilde{M} \in \mathcal{M}$ it holds that*

$$\mathbb{P}_{\tilde{M}} = \mathbb{P}_M \quad \Rightarrow \quad \mathbb{P}^{(X,Y)}_{\tilde{M}(i)} = \mathbb{P}^{(X,Y)}_{M(i)} \quad \forall i \in \mathcal{I},$$

where $\mathbb{P}^{(X,Y)}_{M(i)}$ is the joint distribution of (X, Y) under $M(i)$. Then, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$.

Proposition 3.5 provides verifiable conditions for distribution generalization, and can be used to prove possibility statements. It is, however, not a necessary condition. Indeed, we will see that, under certain types of interventions, distribution generalization becomes possible even in cases where the interventional marginal of X is not identified.

In this section, we study conditions on $\mathcal{M}$, $\mathbb{P}_M$ and $\mathcal{I}$ which ensure generalization, and present corresponding impossibility results proving the necessity of some of these conditions. Two aspects will be of central importance. The first is related to causal identifiability, i.e., whether the causal function f is sufficiently identified from the observational distribution $\mathbb{P}_M$ (Section 3.4.1). The other aspect is related to the types of interventions (Section 3.4.2). We consider interventions on X in Section 3.4.3 and interventions on A in Section 3.4.4. Parts of our results are summarized in Table 3.1.

3.4.1. Identifiability of the Causal Function

For specific types of interventions, the causal function f is itself a minimax solution, see Propositions 3.1 and 3.2. If, in addition, these interventions are support-reducing, generalization is directly implied by the following assumption.

Assumption 3.1 (Identifiability of f on the support of X)**.** *For all $\tilde{M} = (\tilde{f}, \dots) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that $\tilde{f}(x) = f(x)$ for all $x \in \text{supp}(X)$.*

Assumption 3.1 will play a central role in proving distribution generalization even in situations where the causal function is not a minimax solution. We use it as a starting point for most of our results. The assumption is violated, for example, in a linear Gaussian setting with a single covariate X (without A). Here,

in general, we cannot identify f and distribution generalization does not hold. Assumption 3.1, however, is not necessary for generalization. In Section, 3.4.4 we discuss a linear setting where distribution generalization is possible, even if Assumption 3.1 does not hold.

The question of causal identifiability has received a lot of attention in the literature. In linear instrumental variables settings, for example, one assumes that the functions f and g are linear and identifiability follows if the product moment between A and X has rank at least the dimension of X (e.g., Wooldridge, 2010). In linear non-Gaussian models, one can identify the function f even if there are no instruments (Hoyer et al., 2008b). For nonlinear models, restricted SCMs can be exploited, too. In that case, Assumption 3.1 holds under regularity conditions if $h_1(H, \varepsilon_Y)$ is independent of X (Peters et al., 2014, 2017; Zhang and Hyvärinen, 2009) and first attempts have been made to extend such results to non-trivial confounding cases (Janzing et al., 2009). The nonlinear IV setting (e.g., Amemiya, 1974; Newey, 2013; Newey and Powell, 2003) is discussed in more detail in Appendix B.2, where we give a brief overview of identifiability results for linear, parametric and non-parametric function classes. Assumption 3.1 states that f is identifiable, even on $\mathbb{P}_M$-null sets, which is usually achieved by placing further constraints on the function class, such as smoothness. Even though this issue seems technical, it becomes important when considering hard interventions that set X to a fixed value, for example.

3.4.2. Types of Interventions

Whether distribution generalization is admitted depends on the intervention class $\mathcal{I}$. In this work, we only consider interventions on the covariates X and A. Each of these types of interventions can be characterized by a measurable function ψ^i, which determines the structural assignment of the intervened variable, and a (possibly degenerate) random vector I^i, which serves as an independent noise innovation. More formally, for an intervention on X, the pair (ψ^i, I^i) defines the intervention which maps the input model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ to the intervened model $M(i)$ given by the assignments

$$\begin{aligned} A^i &:= \varepsilon_A^i, \quad H^i := \varepsilon_H^i, \\ X^i &:= \psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i), \\ Y^i &:= f(X^i) + h_1(H^i, \varepsilon_Y^i). \end{aligned}$$

Similarly, for an intervention on A, (ψ^i, I^i) specifies the intervention which outputs

$$\begin{aligned} A^i &:= \psi^i(I^i, \varepsilon_A^i), \quad H^i := \varepsilon_H^i, \\ X^i &:= g(A^i) + h_2(H^i, \varepsilon_X^i), \\ Y^i &:= f(X^i) + h_1(H^i, \varepsilon_Y^i). \end{aligned}$$

In both cases, $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$ and $I^i \perp\!\!\!\perp (\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i)$. We will see below that this class of interventions is rather flexible. It does, however, not allow for

arbitrary manipulations of M. For example, it does not allow for changes in the structural assignments for Y or H, or for the noise variable ε_Y^i to enter the assignment of the intervened variable. As the following section highlights, further constraints on the types of interventions are necessary to ensure distribution generalization.

3.4.2.1. Impossibility of Generalization Without Constraints on the Interventions

Let $\mathcal{Q}$ be a class of product distributions on $\mathbb{R}^4$, such that for all $Q \in \mathcal{Q}$, the coordinates of Q are non-degenerate, zero-mean with finite second moment. Let $\mathcal{M}$ be the class of all models of the form

$$A := \varepsilon_A, \quad H := \sigma\varepsilon_H, \quad X := \gamma A + \varepsilon_X + \tfrac{1}{\sigma}H, \quad Y := \beta X + \varepsilon_Y + \tfrac{1}{\sigma}H,$$

with $\gamma, \beta \in \mathbb{R}$, $\sigma > 0$ and $(\varepsilon_A, \varepsilon_X, \varepsilon_Y, \varepsilon_H) \sim Q \in \mathcal{Q}$. Assume that $\mathbb{P}_M$ is induced by some model $M = M(\gamma, \beta, \sigma, Q)$ from the above model class (here, we slightly adapt the notation from Section 3.2). The following proposition shows that, without constraining the set of interventions $\mathcal{I}$, distribution generalization is not always ensured.

Proposition 3.6 (Impossibility of generalization without constraining the class of interventions)**.** *Assume that $\mathcal{M}$ is given as defined above, let $\mathcal{I} \subseteq \mathbb{R}_{>0}$ be a compact, non-empty set and define the interventions on X by $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = iH$, for $i \in \mathcal{I}$. Then, $(\mathbb{P}_M, \mathcal{M})$ does not admit generalization to $\mathcal{I}$ (even if Assumption 3.1 is satisfied). In addition, any prediction model other than the causal model may perform arbitrarily bad under the interventions $\mathcal{I}$. That is, for any $b \neq \beta$ and any $c > 0$, there exists a model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - bX)^2\right] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond X)^2\right] \right| \geq c.$$

We now give some intuition about the above result. By definition, distribution generalization is ensured if there exist prediction functions that are (approximately) minimax optimal for all models which induce the same observational distribution as M. Since, in the above example, the distribution of (X, Y, A) does not depend on σ, this includes all models of the form $M_{\tilde{\sigma}} = M(\gamma, \beta, \tilde{\sigma}, Q)$ for some $\tilde{\sigma} > 0$. However, while agreeing on the observational distribution, each of these models induces fundamentally different intervention distributions (under $M_{\tilde{\sigma}}(i)$, (X, Y) is equal in distribution to $(i\varepsilon_H, (\beta i + \frac{1}{\tilde{\sigma}})\varepsilon_H)$) and results in different (approximate) minimax solutions. Below, we introduce two types of interventions which ensure distribution generalization in a wide range of settings by constraining the influence of H on X.

3.4.2.2. Interventions Which Allow for Generalization

In Section 3.3, we already introduced confounding-removing interventions, which break the dependence between X and H. For an intervention set $\mathcal{I}$ which contains

at least one confounding-removing intervention, the causal function f is always a minimax solution (see Proposition 3.1) and, in the case of support-reducing interventions, distribution generalization is therefore achieved by requiring Assumption 3.1 to hold. The intervention i with intervention map ψ^i is called

$$\begin{gathered}\textit{confounding-preserving} \text{ if there exists a map } \varphi^i, \text{ such that} \\ \psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varphi^i(A^i, g(A^i), h_2(H^i, \varepsilon_X^i), I^i).\end{gathered}$$

Confounding-preserving interventions contain, e.g., *shift interventions* on X, which linearly shift the original assignment by I^i, that is,

$$\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = g(A^i) + h_2(H^i, \varepsilon_X^i) + I^i.$$

The name 'confounding-preserving' stems from the fact that the confounding variables H only enter the intervened structural assignment of X via the term $h_2(H^i, \varepsilon_X^i)$, which is the same as in the original model. (This property fails to hold true for the interventions in Proposition 3.6.) If $\mathcal{I}$ consists only of confounding-preserving interventions, the causal function is generally not a minimax solution. However, we will see that, under Assumption 3.1, these types of interventions lead to identifiability of the intervention distributions $\mathbb{P}_{M(i)}$, $i \in \mathcal{I}$, and therefore ensure generalization via Proposition 3.5.

Some interventions are both confounding-removing and confounding-preserving, but not every confounding-removing intervention is confounding-preserving. For example, the intervention $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varepsilon_X^i$ is confounding-removing but, in general, not confounding-preserving. Similarly, not all confounding-preserving interventions are confounding-removing. We call a set of interventions $\mathcal{I}$ *well-behaved* either if it consists only of confounding-preserving interventions or if it contains at least one confounding-removing intervention.

3.4.3. Generalization to Interventions on X

We now formally prove in which sense the two types of interventions defined above allow for distribution generalization. We will see that this question is closely linked to the relation between the support of $\mathbb{P}_M$ and the support of the intervention distributions. Below, we therefore distinguish between support-reducing and support-extending interventions on X.

3.4.3.1. Support-reducing Interventions

For support-reducing interventions, Assumption 3.1 is sufficient for distribution generalization even in nonlinear settings, under a large class of interventions.

Proposition 3.7 (Generalization to support-reducing interventions on X)**.** *Let $\mathcal{I}$ be a well-behaved set of interventions on X, and assume that* $\operatorname{supp}_{\mathcal{I}}(X) \subseteq \operatorname{supp}(X)$*. Then, under Assumption 3.1, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to the interventions $\mathcal{I}$. If one of the interventions is confounding-removing, then the causal function is a minimax solution.*

In the case of support-extending interventions, further assumptions are required to ensure distribution generalization.

3.4.3.2. Support-extending Interventions

If the interventions in $\mathcal{I}$ extend the support of X, i.e., $\text{supp}_{\mathcal{I}}(X) \not\subseteq \text{supp}(X)$, Assumption 3.1 is not sufficient for ensuring distribution generalization. This is because there may exist a model $\tilde{M} \in \mathcal{M}$ which agrees with M on the observational distribution, but whose corresponding causal function $\tilde{f}$ differs from f outside of the support of X. In that case, a support-extending intervention on X may result in different dependencies between X and Y in the two models, and therefore potentially induce a different set of minimax solutions. The following assumption on the model class $\mathcal{F}$ ensures that any $f \in \mathcal{F}$ is uniquely determined by its values on $\text{supp}(X)$.

Assumption 3.2 (Extrapolation of $\mathcal{F}$)**.** *For all $\tilde{f}, \bar{f} \in \mathcal{F}$ with $\tilde{f}(x) = \bar{f}(x)$ for all $x \in \text{supp}(X)$, it holds that $\tilde{f} \equiv \bar{f}$.*

We will see that this assumption is sufficient (Proposition 3.8) for generalization to well-behaved interventions on X. Furthermore, it is also necessary (Proposition 3.11) if $\mathcal{F}$ is sufficiently flexible. The following proposition can be seen as an extension of Proposition 3.7.

Proposition 3.8 (Generalization to support-extending interventions on X)**.** *Let $\mathcal{I}$ be a well-behaved set of interventions on X. Then, under Assumptions 3.1 and 3.2, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$. If one of the interventions is confounding-removing, then the causal function is a minimax solution.*

Because the interventions may change the marginal distribution of X, the preceding proposition includes examples, in which distribution generalization is possible even if some of the considered joint (test) distributions are arbitrarily far from the training distribution, in terms of any reasonable divergence measure over distributions, such as Wasserstein distance or f-divergence.

Proposition 3.8 relies on Assumption 3.2. Even though this assumption is restrictive, it is satisfied by several reasonable function classes, which therefore allow for generalization to any set of well-behaved interventions. Below, we give two examples of such function classes.

Sufficient conditions for generalization Assumption 3.2 states that every function in $\mathcal{F}$ is globally identified by its values on $\text{supp}(X)$. This is, for example, satisfied if $\mathcal{F}$ is a linear space of functions with domain $\mathcal{D} \subseteq \mathbb{R}^d$ which are linearly independent on $\text{supp}(X)$. More precisely, $\mathcal{F}$ is linearly closed, i.e.,

$$f_1, f_2 \in \mathcal{F}, c \in \mathbb{R}, \implies f_1 + f_2 \in \mathcal{F}, cf_1 \in \mathcal{F}, \tag{3.1}$$

and $\mathcal{F}$ is linearly independent on $\text{supp}(X)$, i.e.,

$$f_1(x) = 0 \quad \forall x \in \text{supp}(X) \implies f_1(x) = 0 \quad \forall x \in \mathcal{D}. \tag{3.2}$$

Examples of such classes include (i) globally linear parametric function classes, i.e., $\mathcal{F}$ is of the form

$$\mathcal{F}^1 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \,|\, \exists \gamma \in \mathbb{R}^k \text{ s.t. } \forall x \in \mathcal{D} : f_\diamond(x) = \gamma^\top \nu(x)\},$$

where $\nu = (\nu_1, \ldots, \nu_k)$ consists of real-valued, linearly independent functions satisfying that $\mathbb{E}_M[\nu(X)\nu(X)^\top]$ is strictly positive definite, and (ii) the class of differentiable functions that extend linearly outside of $\text{supp}(X)$, that is, $\mathcal{F}$ is of the form

$$\mathcal{F}^2 := \left\{ f_\diamond : \mathcal{D} \to \mathbb{R} \;\middle|\; \begin{array}{l} f_\diamond \in C^1 \text{ and } \forall x \in \mathcal{D} \setminus \text{supp}(X) : \\ f_\diamond(x) = f_\diamond(x_b) + \nabla f_\diamond(x_b)(x - x_b) \end{array} \right\}$$

where $x_b := \arg\min_{z \in \text{supp}(X)} \|x - z\|$ and $\text{supp}(X)$ is assumed to be closed with non-empty interior. Clearly, both of the above function classes are linearly closed. To see that $\mathcal{F}^1$ satisfies (3.2), let $\gamma \in \mathbb{R}^k$ be s.t. $\gamma^\top \nu(x) = 0$ for all $x \in \text{supp}(X)$. Then, it follows that $0 = \mathbb{E}_M[(\gamma^\top \nu(X))^2] = \gamma^\top \mathbb{E}_M[\nu(X)\nu(X)^\top]\gamma$ and hence that $\gamma = 0$. To see that $\mathcal{F}^2$ satisfies (3.2), let $f_\diamond \in \mathcal{F}^2$ and assume that $f_\diamond(x) = 0$ for all $x \in \text{supp}(X)$. Then, $f_\diamond(x) = 0$ for all $x \in \mathcal{D}$ and thus $\mathcal{F}^2$ uniquely defines the function on the entire domain $\mathcal{D}$.

By Proposition 3.8, generalization with respect to these model classes is possible for any well-behaved set of interventions. In practice, it may often be more realistic to impose bounds on the higher order derivatives of the functions in $\mathcal{F}$. We now prove that this still allows for what we will call approximate distribution generalization, see Propositions 3.9 and 3.10.

Sufficient conditions for approximate generalization For differentiable functions, exact generalization cannot always be achieved. Bounding the first derivative, however, allows us to achieve approximate generalization. We therefore consider the following function class

$$\mathcal{F}^2 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \,|\, f_\diamond \in C^1 \text{ with } \|\nabla f_\diamond\|_\infty \leq K\} \tag{3.3}$$

for some fixed $K < \infty$, where $\nabla f_\diamond$ denotes the gradient and $\mathcal{D} \subseteq \mathbb{R}^d$. We then have the following result.

Proposition 3.9 (Approx. generalization with bdd. derivatives (confounding-removing))**.** *Let $\mathcal{F}$ be as defined in* (3.3)*. Let $\mathcal{I}$ be a set of interventions on X containing at least one confounding-removing intervention, and assume that Assumption 3.1 holds true. (In this case, the causal function f is a minimax solution.) Then, for all f^* with $f^* = f$ on* $\text{supp}(X)$ *and all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \right| \leq 4\delta^2 K^2 + 4\delta K \sqrt{\text{Var}_M(\xi_Y)},$$

where $\delta := \sup_{x \in \mathrm{supp}_{\mathcal{I}}^{M}(X)} \inf_{z \in \mathrm{supp}^{M}(X)} \|x - z\|$.

If $\mathcal{I}$ *consists only of confounding-removing interventions, the same statement holds when replacing the bound by* $4\delta^2 K^2$.

Proposition 3.9 states that the deviation of the worst-case generalization error from the best possible value is bounded by a term that grows with the square of δ. Intuitively, this means that under the function class defined in (3.3), approximate generalization is reasonable only for interventions that are close to the support of X. We now prove a similar result for cases in which the minimax solution is not necessarily the causal function. The following proposition bounds the worst-case generalization error for arbitrary confounding-preserving interventions. Here, the bound additionally accounts for the approximation to the minimax solution.

Proposition 3.10 (Approx. generalization with bdd. derivatives (confounding-preserving))**.** *Let* $\mathcal{F}$ *be as defined in* (3.3)*. Let* $\mathcal{I}$ *be a set of confounding-preserving interventions on* X*, and assume that Assumption 3.1 is satisfied. Let* $\varepsilon > 0$ *and let* $f^* \in \mathcal{F}$ *be such that,*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] \right| \leq \varepsilon.$$

Then, for all $\tilde{M} \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$*, it holds that*

$$\begin{aligned} &\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \right| \\ &\quad \leq \varepsilon + 12\delta^2 K^2 + 32\delta K \sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K \sqrt{\varepsilon} \end{aligned}$$

where $\delta := \sup_{x \in \mathrm{supp}_{\mathcal{I}}^{M}(X)} \inf_{z \in \mathrm{supp}^{M}(X)} \|x - z\|$.

We can take f^* to be the minimax solution if it exists. In that case, the terms involving ε disappear from the bound, which then becomes more similar to the one in Proposition 3.9.

Impossibility of generalization without constraints on $\mathcal{F}$ If we do not constrain the function class $\mathcal{F}$, generalization is impossible. Even if we consider the set of all continuous functions $\mathcal{F}$, we cannot generalize to interventions outside the support of X. This statement holds even if Assumption 3.1 is satisfied.

Proposition 3.11 (Impossibility of extrapolation)**.** *Assume that* $\mathcal{F} = \{f_\diamond : \mathbb{R}^d \to \mathbb{R} \mid f_\diamond \text{ is continuous}\}$*. Let* $\mathcal{I}$ *be a well-behaved set of support-extending interventions on* X*, such that* $\mathrm{supp}_{\mathcal{I}}(X) \setminus \mathrm{supp}(X)$ *has non-empty interior. Then,* $(\mathbb{P}_M, \mathcal{M})$ *does not admit generalization to* $\mathcal{I}$*, even if Assumption 3.1 is satisfied. In particular, for any function* $\bar{f} \in \mathcal{F}$ *and any* $c > 0$*, there exists a model* $\tilde{M} \in \mathcal{M}$*, with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$*, such that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \bar{f}(X))^2\right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \right| \geq c.$$

The above impossibility result is visualized in Figure 3.1 (left).

3.4.4. Generalization to Interventions on A

We will see that, for interventions on A, parts of the analysis simplify. Since A influences the system only via the covariates X, any such intervention may, in terms of its effect on (X, Y), be equivalently expressed as an intervention on X in which the structural assignment of X is altered in a way that depends on the functional relationship g between X and A. We can therefore employ several of the results from Section 3.4.3 by imposing an additional assumption on the identifiability of g.

Assumption 3.3 (Identifiability of g)**.** *For all $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that $\tilde{g}(a) = g(a)$ for all $a \in \text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$.*

Since $g(A)$ is a conditional mean for X given A, the values of g are identified from $\mathbb{P}_M$ for $\mathbb{P}_M$-almost all a. If $\text{supp}_{\mathcal{I}}(A) \subseteq \text{supp}(A)$, Assumption 3.3 therefore holds if, for example, $\mathcal{G}$ contains continuous functions only. The pointwise identifiability of g is necessary, for example, if some of the test distributions are induced by hard interventions on A, which set A to some fixed value $a \in \mathbb{R}^r$. In the case where the interventions $\mathcal{I}$ extend the support of A, we additionally require the function class $\mathcal{G}$ to extrapolate from $\text{supp}(A)$ to $\text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$; this is similar to the conditions on $\mathcal{F}$ which we made in Section 3.4.3.2 and requires further restrictions on $\mathcal{G}$. Under Assumption 3.3, we obtain a result corresponding to Propositions 3.7 and 3.8.

Proposition 3.12 (Generalization to interventions on A)**.** *Let $\mathcal{I}$ be a set of interventions on A and assume Assumption 3.3 is satisfied. Then, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$ if either $\text{supp}_{\mathcal{I}}(X) \subseteq \text{supp}(X)$ and Assumption 3.1 is satisfied or if both Assumptions 3.1 and 3.2 are satisfied.*

As becomes clear from the proof of this proposition, in general, the causal function does not need to be a minimax solution. Further, Assumption 3.1 is not necessary for generalization. In the case where $\mathcal{F}$, $\mathcal{G}$, $\mathcal{H}_1$ and $\mathcal{H}_2$ consist of linear functions, Rothenhäusler et al. (2021) (anchor regression) and Jakobsen and Peters (2021) (K-class estimators) consider certain sets of interventions on A which render minimax solutions identifiable (and estimate them consistently) even if Assumption 3.1 does not hold. Similarly, if for a categorical A, we have $\text{supp}_{\mathcal{I}}(A) \subseteq \text{supp}(A)$, it is possible to drop Assumption 3.1.

3.4.4.1. Impossibility of Generalization Without Constraining $\mathcal{G}$

Without restrictions on the model class $\mathcal{G}$, generalization to interventions on A is impossible. This holds true even under strong assumptions on the true causal function (such as f is known to be linear). Below, we give a formal impossibility result for hard interventions on A, which set A to some fixed value, and where $\mathcal{G}$ is the set of all continuous functions.

Proposition 3.13 (Impossibility of generalization to interventions on A)**.** *Assume that* $\mathcal{F} = \{f_\diamond : \mathbb{R}^d \to \mathbb{R} \,|\, f_\diamond \text{ is linear}\}$ *and* $\mathcal{G} = \{g_\diamond : \mathbb{R}^r \to \mathbb{R}^d \,|\, g_\diamond \text{ is continuous}\}$. *Let* $\mathcal{A} \subseteq \mathbb{R}^r$ *be bounded, and let* $\mathcal{I}$ *denote the set of all hard interventions which set* A *to some fixed value from* $\mathcal{A}$*. Assume that* $\mathcal{A} \setminus \text{supp}(A)$ *has nonempty interior. Assume further that* $\mathbb{E}_M[\xi_X \xi_Y] \neq 0$ *(this excludes the case of no hidden confounding). Then,* $(\mathbb{P}_M, \mathcal{M})$ *does not admit generalization to* $\mathcal{I}$*. In addition, any function other than* f *may perform arbitrarily bad under the interventions in* $\mathcal{I}$*. That is, for any* $\bar{f} \neq f$ *and* $c > 0$*, there exists a model* $\tilde{M} \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$ *such that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - \bar{f}(X))^2 \right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f_\diamond(X))^2 \right] \right| \geq c.$$

The above impossibility result is visualized in Figure 3.1 (right). This proposition is part of the argument showing that the distribution generalization of anchor regression (Rothenhäusler et al., 2021) can be extended to nonlinear settings only under strong assumptions; the setting of a linear class $\mathcal{G}$ and a potentially nonlinear class $\mathcal{F}$ is covered in Section 3.4.3.2, by rewriting interventions on A as interventions on X.

An impossibility result similar to the proposition above can be shown if A is categorical. As long as not all categories have been observed during training it is possible that the intervention which sets A to a previously unseen category can result in a support-extending distribution shift on X. Using Proposition 3.11, it therefore follows that generalization can become impossible. Since a categorical A can encode settings of multi-task learning and domain generalization (see Section 3.2.4), this result then complements well-known impossibility results for these problems, even under the covariate shift assumption (e.g., Ben-David et al., 2010).

3.5. Learning Generalizing Models from Data

So far, our focus has been on the possibility to generalize, that is, we have investigated under which conditions it is possible to identify generalizing models from the observational distribution. In practice, generalizing models need to be estimated from finitely many data. This task is challenging for several reasons. First, analytical solutions to the minimax problem (3.2) are only known in few cases. Even if generalization is possible, the inferential target thus often remains a complicated object, given as a well-defined but unknown function of the observational distribution. Second, we have seen that the ability to generalize depends strongly on whether the interventions extend the support of X, see Propositions 3.8 and 3.11. In a setting with a finite amount of data, the empirical support of the data lies within some bounded region, and suitable constraints on the function class $\mathcal{F}$ are necessary when aiming to achieve empirical generalization outside this region, even if X comes from a distribution with full support. As we show in our

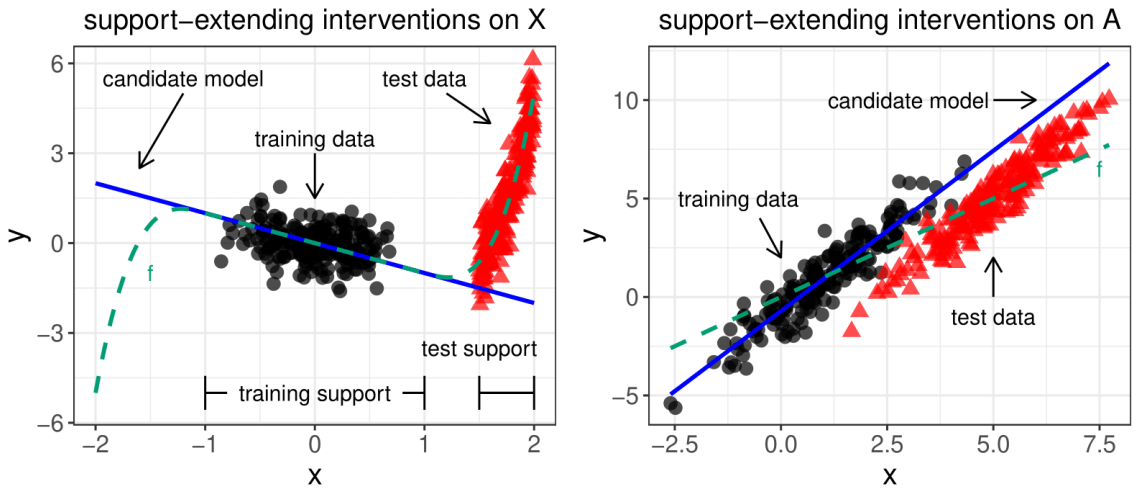


Figure 3.1: Plots illustrating the straight-forward idea behind the impossibility results in Proposition 3.11 (left) and Proposition 3.13 (right). Both plots visualize the case of univariate variables. Under well-behaved interventions on X (left; here using confounding-removing interventions) which extend the support of X, generalization is impossible without further restrictions on the function class $\mathcal{F}$. This holds true even if Assumption 3.1 is satisfied. Indeed, although the candidate model (blue line) coincides with the causal model (green dashed curve) on the support of X, it may perform arbitrarily bad on test data generated under support-extending interventions. Under interventions on A (right) generalization is impossible even under strong assumptions on the function class $\mathcal{F}$ (here, $\mathcal{F}$ is the class of all linear functions). Any support-extending intervention on A shifts the marginal distribution of X by an amount which depends on the (unknown) function g, resulting in a distribution of (X, Y) which, in general, cannot be identified from the observational distribution. Without further restrictions on the function class $\mathcal{G}$, any candidate model apart from the causal model may result in arbitrarily large worst-case risk.

model class	interventions	$\text{supp}_{\mathcal{I}}(X)$	ass.	algorithm
$\mathcal{F}$ linear	on X or A of which at least one is confounding-removing	–	Ass. 3.1	linear IV (e.g., two-stage least squares, K-class or PULSE Jakobsen and Peters (2021); Theil (1958))
$\mathcal{F}, \mathcal{G}$ linear	on A	bounded strength	–	anchor regression Rothenhäusler et al. (2021) and K-class Jakobsen and Peters (2021)
$\mathcal{F}$ smooth	on X or A of which at least one is confounding-removing	support-reducing	Ass. 3.1	nonlinear IV (e.g., NPREGIV Racine and Hayfield (2018), Deep IV (Hartford et al., 2017), Sieve IV (Chen and Christensen, 2018; Newey and Powell, 2003), Kernel IV (Singh et al., 2019))
$\mathcal{F}$ smooth and linearly extrapolates	on X or A of which at least one is confounding-removing	–	Ass. 3.1	**NILE** (Section 3.5.2)

Table 3.2: List of algorithms to learn the generalizing function from data, the considered model class, types of interventions, support under interventions, and additional model assumptions. Sufficient conditions for Assumption 3.1 are given, for example, in the IV literature by generalized rank conditions, see Appendix B.2.

simulations in Section 3.5.2.4 (see figures), constraining the function class can also improve the prediction performance at the boundary of the support.

In Section 3.5.1, we survey existing methods for learning generalizing models. Often, these methods assume either a globally linear model class $\mathcal{F}$ or are completely non-parametric and therefore do not generalize outside the empirical support of the data. Motivated by this observation, we introduce in Section 3.5.2 a novel estimator, which exploits an instrumental variable setup and a particular extrapolation assumption to learn a globally generalizing model.

3.5.1. Existing Methods

As discussed in Section 3.1, a wide range of methods have been proposed to guard against various types of distributional changes. Here, we review methods that fit into the causal framework in the sense that the distributions that in the minimax formulation the supremum is taken over are induced by interventions.

For well-behaved interventions on X which contain at least one confounding-

removing intervention, estimating minimax solutions reduces to the well-studied problem of estimating causal relationships. One class of algorithms for this task is given by linear instrumental variable (IV) approaches. They assume that $\mathcal{F}$ is linear and require identifiability of the causal function (Assumption 3.1) via a rank condition on the observational distribution, see Appendix B.2. Their target of inference is to estimate the causal function, which by Proposition 3.1 will coincide with the minimax solution if the set $\mathcal{I}$ consists of well-behaved interventions with at least one of them being confounding-removing. A basic estimator for linear IV models is the two-stage least squares (TSLS) estimator, which minimizes the norm of the prediction residuals projected onto the subspace spanned by the observed instruments (TSLS objective). TSLS estimators are consistent but do not come with strong finite sample guarantees; e.g., they do not have finite moments in a just-identified setup (e.g., Mariano, 2001). K-class estimators (Theil, 1958) have been proposed to overcome some of these issues. They minimize a linear combination of the residual sum of squares (OLS objective) and the TSLS objective. K-class estimators can be seen as utilizing a bias-variance trade-off. For fixed and non-trivial relative weights, they have, in a Gaussian setting, finite moments up to a certain order that depends on the sample-size and the number of predictors used. If the weights are such that the OLS objective is ignored asymptotically, they consistently estimate the causal parameter (e.g., Mariano, 2001). More recently, PULSE has been proposed (Jakobsen and Peters, 2021), a data-driven procedure for choosing the relative weights such that the prediction residuals 'just' pass a test for simultaneous uncorrelatedness with the instruments.

In cases where the minimax solution does not coincide with the causal function, only few algorithms exist. Anchor regression (Rothenhäusler et al., 2021) is a procedure that can be used when $\mathcal{F}$ and $\mathcal{G}$ are linear and h_1 is additive in the noise component. It finds the minimax solution if the set $\mathcal{I}$ consists of all interventions on A up to a fixed intervention strength, and is applicable even if Assumption 3.1 is not necessarily satisfied.

In a linear setting, where the regression coefficients differ between different environments, it is also possible to minimize the worst-case risk among the observed environments (Meinshausen and Bühlmann, 2015). In its current formulation, this approach does not quite fit into the above framework, as it does not allow for changing distributions of the covariates. A summary of the mentioned methods and their assumptions is given in Table 3.2.

If $\mathcal{F}$ is a nonlinear or non-parametric class of functions, the task of finding minimax solutions becomes more difficult. In cases where the causal function is among such solutions, this problem has been studied in the econometrics community. For example, Newey (2013); Newey and Powell (2003) treat the identifiability and estimation of causal functions in non-parametric function classes. Several non-parametric IV procedures exists, e.g., NPREGIV (Racine and Hayfield, 2018) contains modified implementations of Horowitz (2011) and Darolles et al. (2011), which we will refer to as NPREGIV-1 and NPREGIV-2, respectively. Other procedures include Deep IV (Hartford et al., 2017), Sieve IV (Chen and

Christensen, 2018; Newey and Powell, 2003) and Kernel IV (Singh et al., 2019). Identifiability and estimation of the causal function using nonlinear IV methods in parametric function classes is discussed in Appendix B.2. Unlike in the linear case, most of the methods do not aim to extrapolate and only recover the causal function inside the support of X, that is, they cannot be used to predict interventions outside of this domain. In the following section, we propose a procedure that is able to extrapolate when $\mathcal{F}$ consists of functions which extend linearly outside of the support of X. In principle, any other extrapolation rule may be employed here, as long as all functions from $\mathcal{F}$ are uniquely determined by their values on the support of X, that is, Assumption 3.2 is satisfied.

In our simulations, we see that our method can improve the prediction performance on the boundary of the support and outperforms other methods when comparing the estimation on the support.

3.5.2. NILE

We have seen in Proposition 3.11 that in order to generalize to interventions which extend the support of X, we require additional assumptions on the function class $\mathcal{F}$. In this section, we start from such assumptions and verify both theoretically and practically that they allows us to perform distribution generalization in the considered setup. Along the way, several choices can be made and usually several options are possible. We will see that our choices yield a method with competitive performance, but we do not claim optimality of our procedure. Several of our choices were partially made to keep the theoretical exposition simple and the method computationally efficient. We first consider the univariate case (i.e., X and A are real-valued) and comment later on the possibility to extend the methodology to higher dimensions. Unless specific background knowledge is given, it might be reasonable to assume that the causal function extends linearly outside a fixed interval $[a, b]$. By additionally imposing differentiability on $\mathcal{F}$, any function from $\mathcal{F}$ is uniquely defined by its values within $[a, b]$, see also Section 3.4.3.2. Given an estimate f on $[a, b]$, the linear extrapolation property then yields a global estimate on the whole of $\mathbb{R}$. In principle, any class of differentiable functions can be used. Here, we assume that, on the interval $[a, b]$, the causal function f is contained in the linear span of a B-spline basis. More formally, let $B = (B_1, ..., B_k)$ be a fixed B-spline basis on $[a, b]$, and define $\eta := (a, b, B)$. Our procedure assumes that the true causal function f belongs to the function class $\mathcal{F}_\eta := \{f_\eta(\cdot; \theta) : \theta \in \mathbb{R}^k\}$, where for every $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^k$, $f_\eta(x; \theta)$ is given as

$$f_\eta(x;\theta) := \begin{cases} B(a)^\top\theta + B'(a)^\top\theta(x-a) & \text{if } x < a \\ B(x)^\top\theta & \text{if } x \in [a,b] \\ B(b)^\top\theta + B'(b)^\top\theta(x-b) & \text{if } x > b, \end{cases} \tag{3.1}$$

where $B' := (B'_1, \ldots, B'_k)$ denotes the component-wise derivative of B. In our algorithm, $\eta = (a, b, B)$ is a hyper-parameter, which can be set manually, or be chosen from data.

3.5.2.1. Estimation Procedure

We now introduce our estimation procedure for fixed choices of all hyper-parameters. Section 3.5.2.2 describes how these can be chosen from data in practice. Let $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n\times 3}$ be n i.i.d. realizations sampled from a distribution over (X, Y, A), let $\eta = (a, b, B)$ be fixed and assume that $\text{supp}(X) \subseteq [a, b]$. Our algorithm aims to learn the causal function $f_\eta(\cdot; \theta^0) \in \mathcal{F}_\eta$, which is determined by the linear causal parameter θ^0 of a k-dimensional vector of covariates $(B_1(X), \ldots, B_k(X))$. From standard linear IV theory, it is known that at least k instrumental variables are required to identify the k causal parameters, see Appendix B.2. We therefore artificially generate such instruments by nonlinearly transforming A, by using another B-spline basis $C = (C_1, \ldots, C_k)$. The parameter θ^0 can then be identified from the observational distribution under appropriate rank conditions, see Section 3.5.2.3. In that case, the hypothesis $H_0(\theta) : \theta = \theta^0$ is equivalent to the hypothesis $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top\theta)] = 0$. Let $\mathbf{B} \in \mathbb{R}^{n\times k}$ and $\mathbf{C} \in \mathbb{R}^{n\times k}$ be the associated design matrices, for each $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$ given as $\mathbf{B}_{ij} = B_j(X_i)$ and $\mathbf{C}_{ij} = C_j(A_i)$. A straightforward choice would be to construct the standard TSLS estimator, i.e., $\hat{\theta}$ as the minimizer of $\theta \mapsto \|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, where $\mathbf{P}$ is the projection matrix onto the columns of $\mathbf{C}$; see also Hall (2005). Even though this procedure may result in an asymptotically consistent estimator, there are several reasons why it may be suboptimal in a finite sample setting. First, the above estimator can have large finite sample bias, in particular if k is large. Indeed, in the extreme case where $k = n$, and assuming that all columns in $\mathbf{C}$ are linearly independent, $\mathbf{P}$ is equal to the identity matrix, and $\hat{\theta}$ coincides with the OLS estimator. Second, since θ corresponds to the linear parameter of a spline basis, it seems reasonable to impose constraints on θ which enforce smoothness of the resulting spline function. Both of these points can be addressed by introducing additional penalties into the estimation procedure. Let therefore $\mathbf{K} \in \mathbb{R}^{k\times k}$ and $\mathbf{M} \in \mathbb{R}^{k\times k}$ be the matrices that are, for each $i, j \in \{1, \ldots, k\}$, defined as $\mathbf{K}_{ij} = \int B_i''(x)B_j''(x)dx$ and $\mathbf{M}_{ij} = \int C_i''(a)C_j''(a)da$, and let $\gamma, \delta > 0$ be the respective penalties associated with $\mathbf{K}$ and $\mathbf{M}$. For $\lambda \geq 0$ and with $\mu := (\gamma, \delta, C)$, we then define the estimator

$$\hat{\theta}^n_{\lambda,\eta,\mu} := \underset{\theta\in\mathbb{R}^k}{\arg\min} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma\theta^\top\mathbf{K}\theta, \tag{3.2}$$

where $\mathbf{P}_\delta := \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top$ is the 'hat'-matrix for a penalized regression onto the columns of $\mathbf{C}$. By choice of $\mathbf{K}$, the term $\theta^\top\mathbf{K}\theta$ is equal to the integrated squared curvature of the spline function parametrized by θ. The regularization induced by the second summand in (3.2) is similar to the one from K-class estimators in linear settings (Theil, 1958). The function class (3.1) enforces linear extrapolation. In principle, the above approach extends to situations where X and A are higher-dimensional, in which case B and C consist of multivariate functions. For example, Fahrmeir et al. (2013) propose the use of tensor product splines, and introduce multivariate smoothness penalties based on pairwise first- or second

order parameter differences of basis functions which are close-by with respect to some suitably chosen metric. Similarly to (3.2), such penalties result in a convex optimization problem. However, due to the large number of involved variables, the optimization procedure becomes computationally burdensome already in small dimensions.

Within the function class $\mathcal{F}_\eta$, the above defines the global estimate $f_\eta(x; \hat{\theta}^n_{\lambda,\eta,\mu})$, for every $x \in \mathbb{R}$, given by

$$f_\eta(x; \hat{\theta}^n_{\lambda,\eta,\mu}) := \begin{cases} B(a)^\top \hat{\theta}^n_{\lambda,\eta,\mu} + B'(a)^\top \hat{\theta}^n_{\lambda,\eta,\mu}(x-a) & \text{if } x < a \\ B(x)^\top \hat{\theta}^n_{\lambda,\eta,\mu} & \text{if } x \in [a,b] \\ B(b)^\top \hat{\theta}^n_{\lambda,\eta,\mu} + B'(b)^\top \hat{\theta}^n_{\lambda,\eta,\mu}(x-b) & \text{if } x > b. \end{cases} \tag{3.3}$$

We deliberately distinguish between three different groups of hyper-parameters η, μ and λ. The parameter $\eta = (a, b, B)$ defines the function class to which the causal function f is assumed to belong. To prove consistency of our estimator, we require this function class to be correctly specified. In turn, the parameters λ and $\mu = (\gamma, \delta, C)$ are algorithmic parameters that do not describe the statistical model. Their values only affects the finite sample behavior of our algorithm, whereas consistency is ensured as long as C satisfies certain rank conditions, see Assumption (B2) in Section 3.5.2.3. In practice, γ and δ are chosen via a cross-validation procedure, see Section 3.5.2.2. The parameter λ determines the relative contribution of the OLS and TSLS losses to the objective function. To choose λ from data, we use an idea similar to the PULSE (Jakobsen and Peters, 2021).

3.5.2.2. Algorithm

Let for now η, μ be fixed. In the limit $\lambda \to \infty$, our estimation procedure becomes equivalent to minimizing the TSLS loss $\theta \mapsto \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, which may be interpreted as searching for the parameter θ which complies 'best' with the hypothesis $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top\theta)] = 0$. For finitely many data, following the idea introduced in (Jakobsen and Peters, 2021), we propose to choose the value for λ such that $\tilde{H}_0(\hat{\theta}^n_{\lambda,\eta,\mu})$ is just accepted (e.g., at a significance level $\alpha = 0.05$). That is, among all $\lambda \geq 0$ which result in an estimator that is not rejected as a candidate for the causal parameter, we chose the one which yields maximal contribution of the OLS loss to the objective function. More formally, let for every $\theta \in \mathbb{R}^k$, $T(\theta) = (T_n(\theta))_{n\in\mathbb{N}}$ be a statistical test at (asymptotic) level α for $\tilde{H}_0(\theta)$ with rejection threshold $q(\alpha)$. That is, $T_n(\theta)$ does not reject $\tilde{H}_0(\theta)$ if and only if $T_n(\theta) \leq q(\alpha)$. The penalty $\lambda^\star_n$ is then chosen in the following data-driven way

$$\lambda^\star_n := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) \leq q(\alpha)\}.$$

In general, $\lambda^\star_n$ is not guaranteed to be finite for an arbitrary test statistic T_n. Even for a reasonable test statistic it might happen that $T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) > q(\alpha)$ for all $\lambda \geq 0$; see Jakobsen and Peters (2021) for further details. We can remedy the problem by reverting to another well-defined and consistent estimator, such as the TSLS (which

minimizes the TSLS loss above) if $\lambda_n^\star$ is not finite. Furthermore, if $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda,\eta,\mu})$ is monotonic, $\lambda_n^\star$ can be computed efficiently by a binary search procedure. In our algorithm, the test statistic T and rejection threshold q can be supplied by the user. Conditions on T that are sufficient to yield a consistent estimator $f_\eta(\cdot, \hat{\theta}_{\lambda_n^\star,\mu,\eta})$, given that $\mathcal{F}_\eta$ is correctly specified, are presented in Section 3.5.2.3. Two choices of test statistics which are implemented in our code package can be found in Appendix B.3.

For every $\gamma \geq 0$, let $\mathbf{Q}_\gamma = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \gamma\mathbf{K})^{-1}\mathbf{B}^\top$ be the 'hat'-matrix for the penalized regression onto $\mathbf{B}$. Our algorithm then proceeds as follows.

Algorithm 3.1 NILE ("Nonlinear Intervention-robust Linear Extrapolator")

input: data $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n\times 3}$
options: k, T, q, α
begin
  $a \leftarrow \min_i X_i$, $b \leftarrow \max_i X_i$
  construct cubic B-spline bases $B = (B_1, \ldots, B_k)$ and $C = (C_1, \ldots, C_k)$ at equidistant knots, with boundary knots at respective extreme values of $\mathbf{X}$ and $\mathbf{A}$
  define $\hat{\eta} \leftarrow (a, b, B)$
  choose $\delta^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{P}_\delta\mathbf{Y}$
  choose $\gamma^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma\mathbf{Y}$
  define $\mu^n_{\mathrm{CV}} \leftarrow (\delta^n_{\mathrm{CV}}, \gamma^n_{\mathrm{CV}}, C)$
  approx. $\lambda_n^\star = \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\mu^n_{\mathrm{CV}},\hat{\eta}}) \leq q(\alpha)\}$ by binary search
  update $\gamma^n_{\mathrm{CV}} \leftarrow (1 + \lambda_n^\star) \cdot \gamma^n_{\mathrm{CV}}$
  compute $\hat{\theta}^n_{\lambda_n^\star,\mu^n_{\mathrm{CV}},\hat{\eta}}$ using Equation (3.2)
end
output: $\hat{f}^n_{\mathrm{NILE}} := f_{\hat{\eta}}(\,\cdot\,; \hat{\theta}^n_{\lambda_n^\star,\mu^n_{\mathrm{CV}},\hat{\eta}})$ defined by Equation (3.3)

The penalty parameter γ^n_{CV} is chosen to minimize the out-of-sample mean squared error of the prediction model $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma\mathbf{Y}$, which corresponds to the solution of (3.2) for $\lambda = 0$. After choosing $\lambda_n^\star$, the objective function in (3.2) increases by the term $\lambda_n^\star\|\mathbf{P}_{\delta^n_{\mathrm{CV}}}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In order for the penalty term $\gamma\theta^\top\mathbf{K}\theta$ to impose the same degree of smoothness in the altered optimization problem, the penalty parameter γ needs to be adjusted accordingly. The heuristic update in our algorithm is motivated by the simple observation that for all $\delta, \lambda \geq 0$, $\|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 \leq (1 + \lambda)\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$.

3.5.2.3. Asymptotic Generalization (consistency)

We now prove consistency of our estimator in the case where the hyper-parameters (η, μ) are fixed (rather than data-driven), and the function class $\mathcal{F}_\eta$ is correctly specified. Fix any $a < b$ and a basis $B = (B_1, \ldots, B_k)$. Let $\eta_0 = (a, b, B)$ and let the model class be given by $\mathcal{M} = \mathcal{F}_{\eta_0} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$, where $\mathcal{F}_{\eta_0}$ is as described in Section 3.5.2. Assume that the data-generating model $M = (f_{\eta_0}(\cdot\,; \theta^0), g, h_1, h_2, Q) \in \mathcal{M}$ induces an observational distribution $\mathbb{P}_M$ such that $\text{supp}^M(X) \subseteq (a, b)$. Let further $\mathcal{I}$ be a set of interventions on X or A, and let $\alpha \in (0, 1)$ be a fixed significance level.

We prove asymptotic generalization (consistency) for an idealized version of the NILE estimator which utilizes η_0, rather than the data-driven values. Choose any $\delta, \gamma \geq 0$ and basis $C = (C_1, ..., C_k)$ and let $\mu = (\delta, \gamma, C)$. We will make use of the following assumptions.

(B1) For all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$ it holds that $\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[X^2] < \infty$ and $\sup_{i \in \mathcal{I}} \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top]) < \infty$.

(B2) The product moment matrices $\mathbb{E}_M[B(X)B(X)^\top]$, $\mathbb{E}_M[C(A)C(A)^\top]$, and $\mathbb{E}_M[C(A)B(X)^\top]$ have full rank.

(C1) $T(\theta)$ has uniform asymptotic power on any compact set of alternatives.

(C2) $\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\}$ is almost surely finite.

(C3) $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu})$ is weakly decreasing and $\theta \mapsto T_n(\theta)$ is continuous.

Assumptions (B1)–(B2) ensure consistency of the estimator as long as $\lambda_n^\star$ tends to infinity. Intuitively, in this case, we can apply arguments similar to those that prove consistency of the TSLS estimator. Assumptions (C1)–(C3) ensure that consistency is achieved when choosing $\lambda_n^\star$ in the data-driven fashion described in Section 3.5.2.2. In Assumption (B1), $\lambda_{\max}$ denotes the largest eigenvalue. In words, the assumption states that, under each model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, there exists a finite upper bound on the variance of any linear combination of the basis functions $B(X)$, uniformly over all distributions induced by $\mathcal{I}$. The first two rank conditions of (B2) enable certain limiting arguments to be valid and they guarantee that estimators are asymptotically well-defined. The last rank condition of (B2) is the so-called rank condition for identification. It guarantees that θ^0 is identified from the observational distribution in the sense that the hypothesis $H_0(\theta) : \theta = \theta^0$ becomes equivalent with $\tilde{H}_0(\theta) : \mathbb{E}_M[C(A)(Y - B(X)^\top\theta)] = 0$. (C1) means that for any compact set $K \subseteq \mathbb{R}^k$ with $\theta^0 \notin K$ it holds that $\lim_{n\to\infty} P(\inf_{\theta \in K} T_n(\theta) \leq q(\alpha)) = 0$. If the considered test has, in addition, a level guarantee, such as pointwise asymptotic level, the interpretation of the finite sample estimator discussed in Section 3.5.2.2 remains valid (such level guarantee may potentially yield improved finite sample performance, too). (C2) is made to simplify the consistency proof. As previously

discussed in Section 3.5.2.2, if (C2) is not satisfied, we can output another well-defined and consistent estimator on the event $(\lambda_n^\star = \infty)$, ensuring that consistency still holds.

Under these conditions, we have the following asymptotic generalization guarantee.

Proposition 3.14 (Asymptotic generalization)**.** *Let $\mathcal{I}$ be a set of interventions on X or A of which at least one is confounding-removing. If assumptions (B1)–(B2) and (C1)–(C3) hold true, then, for any $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and any $\varepsilon > 0$, it holds that*

$$\mathbb{P}_M\left(\Big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\Big] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_\diamond(X))^2\Big]\Big| \leq \varepsilon\right)$$

tends to one, as $n \to \infty$. In the above event, only $\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}$ is stochastic.

3.5.2.4. Experiments

We now investigate the empirical performance of our proposed estimator, the NILE, with $k = 50$ spline basis functions. To choose $\lambda_n^\star$, we use the test statistic T_n^2, which tests the slightly stronger hypothesis $\bar{H}_0$, see Appendix B.3. In all experiments use the significance level $\alpha = 0.05$. We include two other approaches as baseline: (i) the method NPREGIV-1 (using its default options) introduced in Section 3.5.1, and (ii) a linearly extrapolating estimator of the ordinary regression of Y on X (which corresponds to the NILE with $\lambda^\star \equiv 0$). In all experiments, we generate data sets of size $n = 200$ as independent replications from

$$\begin{aligned} &A := \varepsilon_A, \quad H := \varepsilon_H, \quad X := \alpha_A A + \alpha_H H + \alpha_\varepsilon \varepsilon_X, \\ &Y := f(X) + 0.3H + 0.2\varepsilon_Y, \end{aligned} \tag{3.4}$$

where $(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ are jointly independent with Uniform$(-1, 1)$ marginals. To make results comparable across different parameter settings, we impose the constraint $\alpha_A^2 + \alpha_H^2 + \alpha_\varepsilon^2 = 1$, which ensures that in all models, X has variance $1/3$. The function f is drawn from the linear span of a basis of four natural cubic splines with knots placed equidistantly within the 90% inner quantile range of X. By well-known properties of natural splines, any such function extends linearly outside the boundary knots. Figure 3.2 (left) shows an example data set from (3.4), where the causal function is indicated in green. We additionally display estimates obtained by each of the considered methods, based on 20 i.i.d. datasets. Due to the confounding variable H, the OLS estimator is clearly biased. NPREGIV-1 exploits A as an instrumental variable and obtains good results within the support of the observed data. Due to its non-parametric nature, however, it cannot extrapolate outside this domain. The NILE estimator exploits the linear extrapolation assumption on f to produce global estimates.

We further investigate the empirical worst-case risk across several different models of the form (3.4). That is, for a fixed set of parameters $(\alpha_A, \alpha_H, \alpha_\varepsilon)$, we

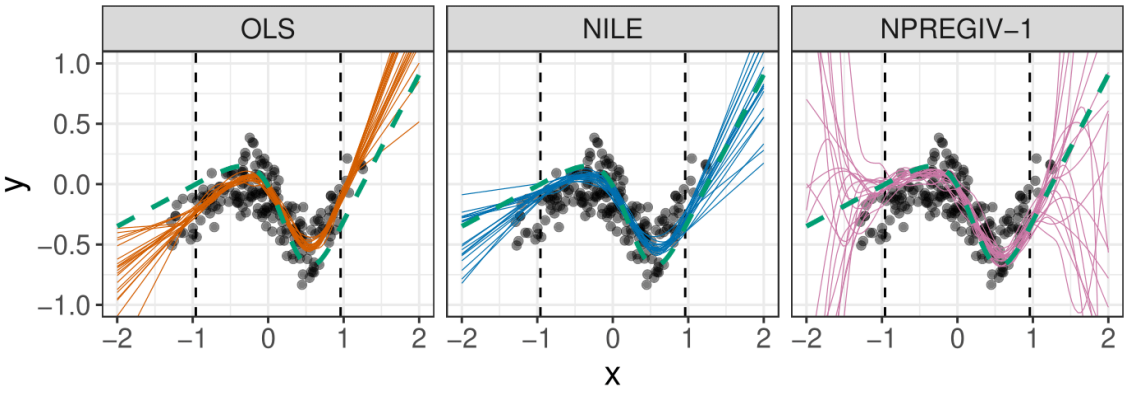


Figure 3.2: A sample dataset from the model (3.4) with $\alpha_A = \sqrt{1/3}$, $\alpha_H = \sqrt{2/3}$, $\alpha_\varepsilon = 0$. The true causal function is indicated by a green dashed line. For each method, we show 20 estimates of this function, each based on an independent sample from (3.4). For values within the support of the training data (vertical dashed lines mark the inner 90% quantile range), NPREGIV-1 correctly estimates the causal function well. As expected, when moving outside the support of X, the estimates become unreliable, and we gain an increasing advantage by exploiting the linear extrapolation assumed by the NILE.

construct several models $M_1, \ldots, M_N$ of the form (3.4) by randomly sampling causal functions $f_1, \ldots, f_N$ (see Appendix B.4 for further details on the sampling procedure). For every $x \in [0, 2]$, let $\mathcal{I}_x$ denote the set of hard interventions which set X to some fixed value in $[-x, x]$. We then characterize the performance of each method using the average (across different models) worst-case risk (across the interventions in $\mathcal{I}_x$), i.e., for each estimator $\hat{f}$, we consider

$$\frac{1}{N}\sum_{j=1}^{N} \sup_{i \in \mathcal{I}_x} \mathbb{E}_{M_j(i)}\left[(Y - \hat{f}(X))^2\right] = \mathbb{E}[\xi_Y^2] + \frac{1}{N}\sum_{j=1}^{N} \sup_{\tilde{x} \in [-x,x]} (f_j(\tilde{x}) - \hat{f}(\tilde{x}))^2, \quad (3.5)$$

where $\xi_Y := 0.3H + 0.2\varepsilon_Y$ is the noise term for Y (which is fixed across all experiments). In practice, we evaluate the functions $\hat{f}$, $f_1, \ldots, f_N$ on a fine grid on $[-x, x]$ to approximate the above supremum. Figure 3.3 plots the average worst-case risk versus intervention strength for varying degree of confounding (α_H). The optimal worst-case risk $\mathbb{E}[\xi_Y^2]$ is indicated by a green dashed line. The results show that the linear extrapolation property of the NILE estimator is beneficial in particular for strong interventions. In the case of no confounding ($\alpha_H = 0$), the minimax solution coincides with the regression of Y on X, hence even the OLS estimator yields good predictive performance. In this case, the hypothesis $\bar{H}_0(\hat{\theta}^n_{\lambda,\delta^n_{\mathrm{CV}},\gamma^n_{\mathrm{CV}}})$ is accepted already for small values of λ (in this experiment, the empirical average of $\lambda^\star_n$ equals 0.015), and the NILE estimator becomes indistinguishable from the OLS. As the confounding strength increases, the OLS becomes increasingly biased, and the

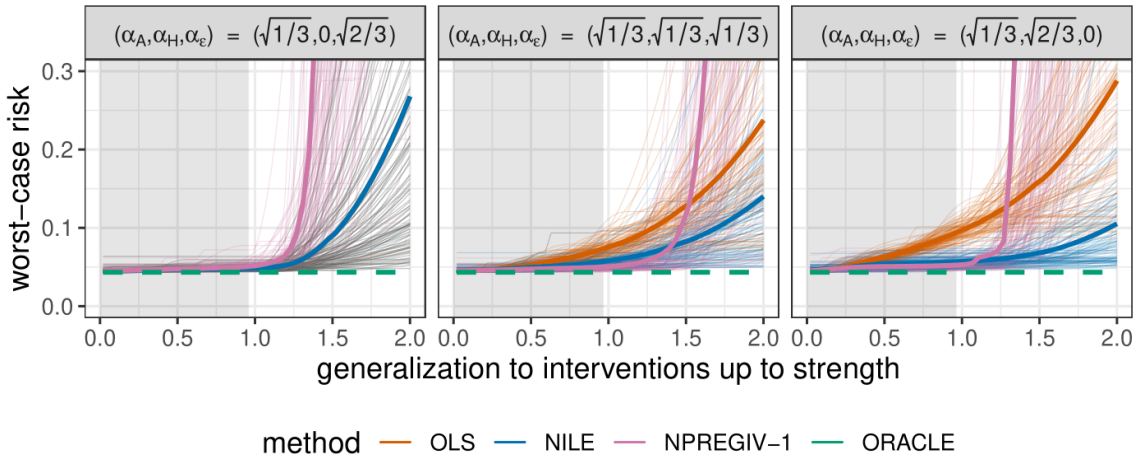


Figure 3.3: Predictive performance under confounding-removing interventions on X for different confounding- and intervention strengths (see alpha values in the grey panel on top). The right panel corresponds to the same parameter setting as in Figure 3.2. The plots in each panel are based on data sets of size $n = 200$, generated from $N = 100$ different models of the form (3.4). For each model, we draw a different function f, resulting in a different minimax solution (see Appendix B.4 for details on the sampling procedure). The performances under individual models are shown by thin lines; the average performance (3.5) across all models is indicated by thick lines. In all considered models, the optimal prediction error (green dashed line) is equal to $\mathbb{E}[\xi_Y^2]$ (by consistency, for any fixed function f, NILE's worst-case risk converges pointwise to this value for increasing sample size). The grey area indicates the inner 90 % quantile range of X in the training distribution; the white area can be seen as an area of generalization.

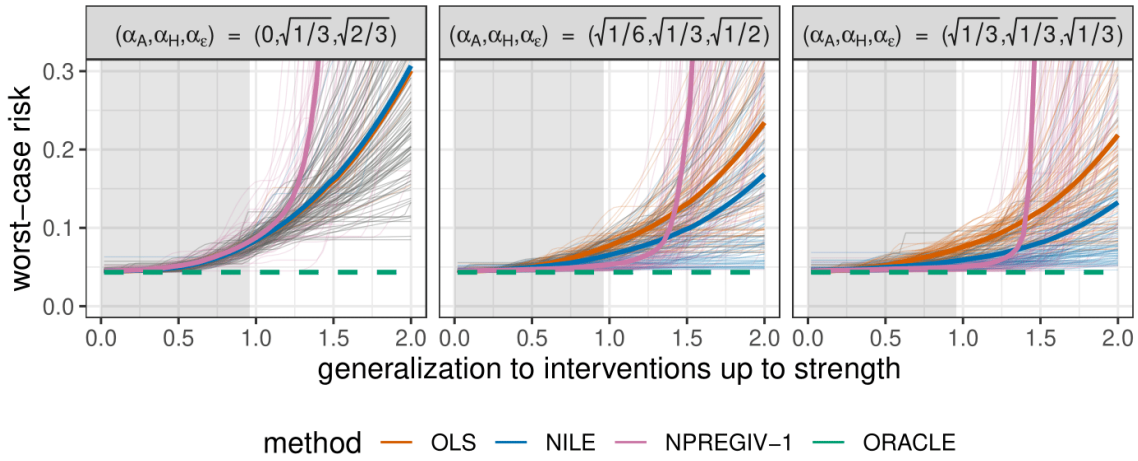


Figure 3.4: Predictive performance for varying instrument strength. If the instruments have no influence on X ($\alpha_A = 0$), the second term in the objective function (3.2) is effectively constant in θ, and the NILE therefore coincides with the OLS estimator (which uses $\lambda = 0$). This guards the NILE against the large variance which most IV estimators suffer from in a weak instrument setting. For increasing influence of A, it clearly outperforms both alternative methods for large intervention strengths.

NILE objective function differs more notably from the OLS (average $\lambda_n^\star$ of 2.412 and 5.136, respectively). The method NPREGIV-1 slightly outperforms the NILE inside the support of the observed data, but drops in performance for stronger interventions. We believe that the increase in extrapolation performance of the NILE for stronger confounding (increasing α_H) might stem from the fact that, as the $\lambda_n^\star$ increases, also the smoothness penalty γ increases, see Algorithm 3.1. While this results in slightly worse in-sample prediction, it seems beneficial for extrapolation (at least for the particular function class that we consider). We do not claim that our algorithm has theoretical guarantees which explain this increase in performance.

Figure 3.4 shows the worst-case risk for varying instrument strength (α_A). In the case where all exogenous noise comes from the unobserved variable ε_X (i.e., $\alpha_A = 0$), the NILE coincides with the OLS estimator. In such settings, standard IV methods are known to perform poorly, although also the NPREGIV-1 method seems robust to such scenarios. As the instrument strength increases, the NILE clearly outperforms OLS and NPREGIV-1 for interventions on X which include values outside the training data.

We further compare NILE's ability to estimate the causal function on the support of the covariate X in a nonlinear IV setting and compare it with the results from other state-of-the-art procedures for nonlinear IV estimation, following the experimental setup by Singh et al. (2019). Here, the authors consider a predictor

variable $X \sim \text{Uniform}(0, 1)$ which causally influences the target variable Y via the structural assignment $Y := f(X) + \xi_Y$, where f is the nonlinear causal function $f(x) = \log(|16x - 8| + 1) \cdot \text{sgn}(x - 1/2)$, and ξ_Y is an additive error term which is correlated with X. They compare their proposed procedure Kernel IV to the methods NPREGIV-2 (Singh et al. (2019) refer to this method as 'Smooth IV'), Sieve IV and Deep IV (see Section 3.5.1). As a baseline, they also include a method for standard kernel ridge regression ('Kernel Reg') (Saunders et al., 1998), which ignores the existence of hidden confounders. Each procedure yields a different estimator $\hat{f}$. Based on 40 independent simulations, the estimators are then compared in terms of the average squared distance between f and $\hat{f}$ across 1000 equidistant points in the interval $[0, 1]$. We refer to (Singh et al., 2019, Appendix A.11) for a precise description of the experimental setup. Figure 3.5 shows the results of the above experiment (corresponding to Figure 2 in (Singh et al., 2019)), where we have also included the NILE. Our method outperforms all other procedures, in particular for large sample sizes. There is slight difference in the way the different algorithms use the available data. In order to reduce finite sample bias, Singh et al. (2019) use sample splitting, where the first and second step of the two-stage-least-squares procedure are performed on disjoint data sets. The NILE, in contrast, uses all of the data at once. However, even when running our procedure on only half of the data, we still outperform the other procedures by a distinct margin, see Figure B.3. We believe that the superior MSE performance of NILE could be due to the different approaches of regularization. For example, NILE uses causal regularization similar to that of PULSE, i.e., a data-driven K-class regularization; in linear IV settings, this type of regularization often yields a smaller MSE than standard IV methods such as TSLS (Jakobsen and Peters, 2021).

3.6. Discussion and Future Work

In many real world problems, the test distribution may differ from the training distribution. This requires statistical methods that come with a provable guarantee in such a setting. It is possible to characterize robustness by considering predictive performance for distributions that are close to the training distribution in terms of standard divergences or metrics, such as KL divergences or Wasserstein distance. As an alternative view point, we have introduced a novel framework that formalizes the task of distribution generalization when considering distributions that are induced by a set of interventions. Based on the concept of modularity, interventions modify parts of the joint distribution and leave other parts invariant. Thereby, they impose constraints on the changes of the distributions that are qualitatively different from considering balls in the above metrics. As such, we see them as a useful language to describe realistic changes between training and test distributions.

Our framework is general in that it allows us to model a wide range of causal models and interventions, which do not need to be known beforehand. We

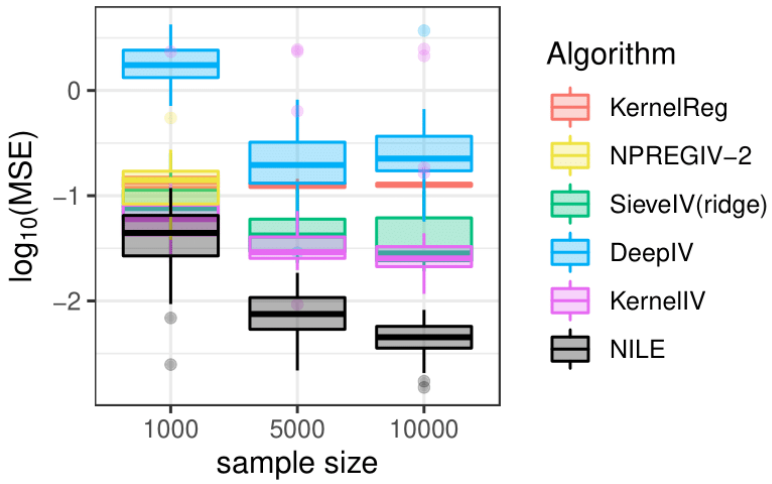


Figure 3.5: Comparison between the NILE and several alternative procedures for learning a nonlinear causal function, based on the same experimental setup as in Singh et al. (2019). The estimated functions are evaluated on the support (no generalization). NILE outperforms the competing methods.

have proved several generalization guarantees, some of which show robustness for distributions that are not close to the training distribution by considering almost any of the standard metrics. Here, generalization can be obtained by causal functions, but also by non-causal functions; in general, however, the minimizer changes when the intervention class is altered (or misspecified). We have further proved impossibility results that indicate the limits of what is possible to learn from the training distribution. In particular, in nonlinear models, strong assumptions are required for distribution generalization to a different support of the covariates. As such, methods such as anchor regression cannot be expected to work in nonlinear models, unless strong restrictions are placed on the function class $\mathcal{G}$.

Our work can be extended into several directions. It may, for example, be worthwhile to investigate the sharpness of the bounds we provide in Section 3.4.3.2 and other extrapolation assumptions on $\mathcal{F}$. Our results make use of the form of the squared loss and it remains an open question to which extent they hold for general convex loss functions. While our results can be applied to situations where causal background knowledge is available, via a transformation of SCMs, our analysis is deliberately agnostic about such information. It would be interesting to see whether stronger theoretical results can be obtained by including causal background information. We showed that the type of the interventions play a crucial role in determining whether the causal function is a minimax optimal solution. Building on this, it would be interesting to find empirical procedures which test whether an intervention is confounding-removing, confounding-preserving or neither. Finally, it could be worthwhile to investigate whether NILE, which outperforms existing approaches with respect to extrapolation, can be combined with non-parametric

methods to further improve in-sample performance. While our current framework already contains certain settings of multi-task learning and domain generalization, it could be instructive to additionally include the possibility to model unlabeled data in the test task. Finally, our results concern the infinite sample case, but we believe that they can form the basis for a corresponding analysis involving rates or even finite sample results.

We view our work as a step towards understanding the problem of distribution generalization. We hope that considering the concepts of interventions may help to shed further light into the question of generalizing knowledge that was acquired during training to a different test distribution.

Acknowledgments

We thank Thomas Kneib for helpful discussions and two anonymous reviewers for valuable comments. RC and JP were supported by a research grant (18968) from VILLUM FONDEN; MEJ and JP were supported by the Carlsberg Foundation.

Chapter 4

Structure Learning for Directed Trees

Joint work with

Rajen Shah, Peter Bühlmann and Jonas Peters

Abstract

Knowing the causal structure of a system is of fundamental interest in many areas of science and can aid the design of prediction algorithms that work well under manipulations to the system. The causal structure becomes identifiable from the observational distribution under certain restrictions. To learn the structure from data, score-based methods evaluate different graphs according to the quality of their fits. However, for large nonlinear models, these rely on heuristic optimization approaches with no general guarantees of recovering the true causal structure. In this paper, we consider structure learning of directed trees. We propose a fast and scalable method based on Chu–Liu–Edmonds' algorithm we call causal additive trees (CAT). For the case of Gaussian errors, we prove consistency in an asymptotic regime with a vanishing identifiability gap. We also introduce a method for testing substructure hypotheses with asymptotic family-wise error rate control that is valid post-selection and in unidentified settings. Furthermore, we study the identifiability gap, which quantifies how much better the true causal model fits the observational distribution, and prove that it is lower bounded by local properties of the causal model. Simulation studies demonstrate the favorable performance of CAT compared to competing structure learning methods.

Keywords: Causality, restricted causal models, structure learning, directed trees, hypothesis testing.

4.1. Introduction

Learning the underlying causal structure of a stochastic system involving the random vector $X = (X_1, \dots, X_p)$ is an important problem in economics, industry, and science. Knowing the causal structure allows researchers to understand whether X_i causes X_j (or vice versa) and how a system reacts under an intervention.

However, it is not generally possible to learn the causal structure (or parts thereof) from the observational data of a system alone. Without further restrictions on the system of interest there might exist another system with a different causal structure inducing the same observational distribution, i.e., the structure might not be identifiable from observed data.

Common structure learning methods using observational data are constraint-based (e.g., Pearl, 2009; Spirtes et al., 2000), score-based (e.g., Chickering, 2002), or a mix thereof (e.g., Nandy et al., 2018). Each of these approaches requires different assumptions to ensure identifiability of the causal structure and consistency of the approach. In structural causal models, one assumes that there are (causal) functions $f_1, \ldots, f_p$ such that for all

$$1 \leq i \leq p: \qquad X_i := f_i(X_{\mathrm{PA}(i)}, N_i),$$

for subsets $\mathrm{PA}(i) \subseteq \{1, ..., p\}$ and jointly independent noise variables $N = (N_1, ..., N_p) \sim P_N$ (see Definition 4.1 for a precise definition including further restrictions). The causal graph is constructed as follows: for each variable X_i one adds directed edges from its direct causes or parents $\mathrm{PA}(i)$ into i. For such models, system assumptions concerning the causal functions can make the causal graph identified from the observational distribution. Specific assumptions that guarantee identifiability of the causal graph have been studied for, e.g., linear Gaussian models with equal noise variance (Peters and Bühlmann, 2014), linear non-Gaussian models (Shimizu et al., 2006), nonlinear additive noise models (Hoyer et al., 2008a; Peters et al., 2014), partially-linear additive Gaussian models (Rothenhäusler et al., 2018) and discrete models (Peters et al., 2011).

Score-based structure learning usually starts with a function ℓ assigning a population score to causal structures. Depending on the assumed model class, this function is minimized by the true structure. For example, when considering directed acylic graph (DAGs), the true causal DAG $\mathcal{G}$ satisfy

$$\mathcal{G} = \underset{\tilde{\mathcal{G}}\,:\,\tilde{\mathcal{G}} \text{ is a DAG}}{\arg\min} \ \ell(\tilde{\mathcal{G}}). \tag{4.1}$$

The idea is then to estimate the score from a finite sample and minimize the empirical score over all DAGs. As the cardinality of the space of all DAGs grows super-exponentially in the number of nodes p (Chickering, 2002), brute-force minimization becomes computationally infeasible even for moderately large systems.[1]

For linear Gaussian models, assuming the Markov conditions and faithfulness, one can recover the correct Markov equivalence class (MEC) of $\mathcal{G}$, which can be represented by a unique completed partially directed acyclic graph (CPDAG) (Pearl, 2009). The optimization can be done greedily over MECs or DAGs (Chickering, 2002; Tsamardinos et al., 2006) and in the former case, the method is known to be consistent (Chickering, 2002). In the nonlinear case, Bühlmann et al. (2014)

[1] For example, there are over 10^{275} distinct directed acyclic graphs over 40 nodes (Sloane, 2021).

show that nonparametric maximum-likelihood estimation consistently estimates the correct causal order. However, the greedy search algorithm minimizing the score function does not come with any theoretical guarantees. Recently, methods have been proposed that perform continuous, non-convex optimization (Zheng et al., 2018) but such methods are without guarantees and it is currently debated whether they exploit some artifacts in simulated data (Reisach et al., 2021). Thus, for nonlinear models, there is currently no score-based method that guarantees recovery of the true causal graph with high probability.

This paper focuses on models of reduced complexity, namely models with directed trees as causal graphs. We will show that this complexity reduction allow for computationally feasible minimization of the score-function using the Chu–Liu–Edmonds' algorithm (proposed independently by Chu and Liu, 1965; Edmonds, 1967). Our method is called causal additive trees (CAT). The method is easy to implement and consists of two steps. In the first step, we employ user-specified (univariate) regression methods to estimate the pairwise conditional means of each variable given all other variables. We then use these to construct edge weights as inputs to the Chu–Liu–Edmonds' algorithm. This algorithm then outputs a directed tree with minimal edge weight, corresponding to a directed tree minimizing the score in Equation (4.1).

4.1.1. Contributions

We now highlight four main contributions of the paper:

(i) Computational feasibility: Assuming an identifiable model class, such as additive noise, allows us to infer the causal DAG by minimizing Equation (4.1) for a suitable score function. However, even for trees, the cardinality of the search space grows super-exponentially in the number of variables p. Hence, brute-force minimization (exhaustive search) in Equation (4.1) remains computationally infeasible for large systems. We propose the score-based method CAT and prove that it recovers the causal tree with a run-time complexity of $\mathcal{O}(p^2)$.

(ii) Consistency: We prove that CAT is pointwise consistent in an identified Gaussian noise setup. That is, we recover the causal directed tree with probability tending to one as the sample size increases. Consistency only requires that the regression methods for estimating the conditional mean functions have mean squared prediction error converging to zero in probability. This property that is satisfied by many nonparametric regression methods such as nearest neighbors, neural networks, or kernel methods (see e.g. Györfi et al., 2002). Moreover, the vanishing estimation error is only required for causal edges for which the conditional means coincide with the causal functions. We also derive sufficient conditions that ensure consistency in an asymptotic setup with vanishing identifiability. Specifically, we show that consistency is retained even when the identifiability gap decreases at a rate q_n with $q_n^{-1} = o(\sqrt{n})$ as long as the conditional expectation mean squared prediction error corresponding to the causal edges vanishes at a rate $o_p(q_n)$.

(iii) Hypothesis testing: We provide an algorithm for performing hypothesis tests concerning the presence and absence of substructures, such as particular edges, in the true causal graph. The type I error is controlled asymptotically when the mean squared prediction error of the regression corresponding to the true causal edges decays at a relatively slow $o_p(n^{-1/2})$ rate. The tests are valid post-selection, that is, the hypotheses to be tested may be chosen after the graph has been estimated, and when multiple tests are performed, the family-wise error rate is controlled for any number of tests. In the non-identified setting where multiple minimizers of the population score exist, the inferences derived are valid for the set of minimizers, so one can for instance test whether a particular edge is present in all graphs minimizing the score.

(iv) Identifiability analysis: We analyze the identifiability gap, that is, the smallest population score difference between an alternative graph and the causal graph. The reduced system complexity, due to the restriction to trees, allows us to derive simple yet informative lower bounds. For Gaussian additive models, for example, the lower bound can be computed using only local properties of the underlying model: it is based on a first term that considers the minimal score gap between individual edge reversals and a second term involving the minimal mutual information of two neighboring nodes, when conditioning on another neighbor of the parent node.

4.1.2. Related Constraint-based Approaches

As an alternative to score-based methods, constraint-based methods such as PC or FCI (Spirtes et al., 2000) test for conditional independences statements in P_X and use these results to infer (parts of) the causal structure. Such methods usually assume that P_X is both Markov and faithful with respect to the causal graph $\mathcal{G}$. Under these assumptions, the Markov equivalence class of the causal graph $\mathcal{G}$ is identified. In a jointly Gaussian setting, consistency of constraint-based approaches relies on faithfulness, whereas uniform consistency requires strong faithfulness (see, e.g., Kalisch and Bühlman, 2007; Zhang and Spirtes, 2002) – a condition that has been shown to be strong (Uhler et al., 2013). In nonlinear settings, corresponding guarantees do not exist. This may at least partially be due to the fact that conditional independence testing is known to be a hard statistical problem (Shah and Peters, 2020).

Constraint-based methods have also been studied for polytrees. A polytree is a DAG whose undirected graph is a tree. Polytrees, unlike directed trees, allow for multiple root nodes as well as nodes with multiple parents. Rebane and Pearl (1987), inspired by the work of Chow and Liu (1968), propose a constraint-based structure learning method for polytrees over discrete variables that can identify the correct skeleton and causal basins, structures constructed from nodes with at least two parents. More precisely, the skeleton is determined by the maximum weight spanning tree (MWST) algorithm with mutual information measure weights, while the directionality of edges is inferred by conditional independence constraints

implied by the observed distribution. In the case of causal trees this constraint-based structure learning method cannot direct any edges because causal basins do not exist (Rebane and Pearl, 1987). Dominguez et al. (2013) and Ouerd (2000) extend the Rebane and Pearl (1987) algorithm for causal discovery to multivariate Gaussian polytree distributions. In this work, we employ Chu–Liu–Edmonds' algorithm, a directed analogue of the MWST algorithm, to not only recover the skeleton but also the direction of all edges in the causal graph. This is possible since we consider restricted causal models, e.g., nonlinear additive Gaussian noise models. (When discarding information that allows us to infer directionality of the edges, one recovers the mutual information weights of Rebane and Pearl (1987), see Remark C.1 in Appendix C.2 for details.)

4.1.3. Organization of the Paper

In Section 4.2, we define the setup and relevant score functions. We further strengthen existing identifiability results for nonlinear additive noise models. In Section 4.3, we propose CAT, an algorithm solving the score-based structure learning problem that is based on Chu–Liu–Edmonds' algorithm. We prove consistency of CAT for a fixed distribution and for a setup with vanishing identifiability. In Section 4.4, we provide results on asymptotic normality of the scores, construct confidence regions and propose feasible testing procedures. Section 4.5, we analyzes the identifiability gap. Section 4.6 shows the results of various simulation experiments. All proofs can be found in Appendix C.4.

4.2. Score-based Learning and Identifiability of Trees

In the remainder of this work we use of the following graph terminology (a more detailed introduction can be found in Appendix C.1, see also Koller and Friedman, 2009). A directed graph $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (or nodes) $V = \{1, \dots, p\}$ and a collection of directed edges $\mathcal{E} \subseteq \{(i \to j) \equiv (i, j) : i, j \in V, i \neq j\}$. A directed acyclic graph (DAG) is a directed graph that does not contain any directed cycles. A directed tree is a connected DAG in which all nodes have at most one parent. The unique node of a directed tree $\mathcal{G}$ with no parents is called the root node and is denoted by $\mathrm{rt}(\mathcal{G})$. We let $\mathcal{T}_p$ denote the set of directed trees over $p \in \mathbb{N}_{>0}$ nodes.

4.2.1. Identifiability of Causal Additive Tree Models

We now revisit and strengthen known identifiability results on restricted structural causal models. Consider a distribution that is induced by a structural causal model (SCM) with additive noise. Then, there are only special cases (such as linear Gaussian models) for which alternative models with a different causal structure exist that generate the same distribution (see Peters et al., 2017, for an overview). To state and strengthen these results formally, we introduce the following notation.

For any $k \in \mathbb{N}$ we define the following classes of functions from $\mathbb{R}$ to $\mathbb{R}$: $\mathcal{M}$ denotes all measurable functions, $\mathcal{D}_k$ denotes the set of all k times differentiable functions and $\mathcal{C}_k$ denotes the k times continuously differentiable functions. We let $\mathcal{P}$ denote the set of mean zero probability measures on $\mathbb{R}$ that have a density with respect to Lebesgue measure. $\mathcal{P}_+ \subseteq \mathcal{P}$ denotes the subset for which a density is strictly positive. For any function class $\mathcal{F} \subseteq \{f | f : \mathbb{R} \to \mathbb{R}\}$, $\mathcal{P}_\mathcal{F} \subseteq \mathcal{P}$ denotes the subset with a density function in $\mathcal{F}$. As a special case, we let $\mathcal{P}_\mathrm{G} \subseteq \mathcal{P}_{+\mathcal{C}_\infty} := \mathcal{P}_+ \cap \mathcal{P}_{\mathcal{C}_\infty}$ denote the subset of Gaussian probability measures. For any set $\mathcal{P}$ of probability measures, $\mathcal{P}^p$ denotes all p-dimensional product measures on $\mathbb{R}^p$ with marginals in $\mathcal{P}$.

We now define structural causal additive tree models as SCMs with a tree structure.

Definition 4.1 (Structural causal additive tree models)**.** *Consider a class $\mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$. Any tuple $(\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$ induces a structural causal model over $X = (X_1, \ldots, X_p)$ given by the following structural assignments*

$$X_i := f_i(X_{\mathrm{pa}^\mathcal{G}(i)}) + N_i, \quad \text{for all } 1 \leq i \leq p,$$

where $f_{\mathrm{rt}(\mathcal{G})} \equiv 0$ and $N = (N_1, \ldots, N_p) \sim P_N$, which we call a structural causal additive tree model. By slight abuse of notation, we write $Q \in \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$ for a probability distribution that is induced by a structural causal additive tree model.

Furthermore, we define the set of restricted structural causal additive tree models. We will see later that for these models, the causal graph is identifiable from the observable distribution of the system. When the causal graph of a sufficiently nice additive noise SCM is not identifiable, then certain differential equations must hold (see the proof of Proposition 4.1 for details). The definition of restricted structural causal additive tree models ensures that this does not happen.

Definition 4.2 (Restricted structural causal additive tree models)**.** *The collection of restricted structural causal additive tree models (or causal additive tree models, for short) $\Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ is given by all models $\theta = (\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ satisfying the following conditions for all $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$: (i) $f_i \in \mathcal{D}_3$, (ii) f_i is nowhere constant, i.e., it is not constant on any open set, and (iii) the induced log-density ξ of $X_{\mathrm{pa}^\mathcal{G}(i)}$, noise log-density ν of N_i and causal function f_i are such that for all $x, y \in \mathbb{R}$ such that $\nu''(y - f_i(x))f_i'(x) \neq 0$ it holds that*

$$\xi''' \neq \xi'' \left(\frac{f_i''}{f_i'} - \frac{\nu''' f_i'}{\nu''} \right) - 2\nu'' f_i'' f_i' + \nu' f_i''' + \frac{\nu' \nu''' f_i'' f_i'}{\nu''} - \frac{\nu' (f_i''')^2}{f'}, \tag{4.1}$$

where the derivatives of ξ, ν and f_i are evaluated in x, $y - f_i(x)$ and x, respectively.

The following lemma, due to Hoyer et al. (2008a), shows that for additive Gaussian noise models, the differential equation constraints of Definition 4.2 simplify.[2] We obtain identifiability (by Proposition 4.1) if the causal functions are nonlinear.

[2]For completeness, we include the proof of Lemma 4.1 in Appendix C.4, using the approach of Zhang and Hyvärinen (2009) but expressed in our notation.

Lemma 4.1. *Let $\theta = (\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$. Assume that for all $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$ the following three conditions hold (a) $f_i \in \mathcal{D}_3$, (b) f_i is nowhere constant and (c) f_i is not linear. Then, $\theta \in \Theta_R$.*

Existing identifiability results for causal graphs in restricted SCMs (Hoyer et al., 2008a; Peters et al., 2014) are stated and proven in terms of the ability to distinguish the induced distributions of two restricted structural causal models: For all $\theta = (\mathcal{G}, \ldots) \in \Theta_R$ and $\tilde{\theta} = (\tilde{\mathcal{G}}, \ldots) \in \Theta_R$, if $\mathcal{G} \neq \tilde{\mathcal{G}}$, then $\mathcal{L}(X_\theta) \neq \mathcal{L}(X_{\tilde{\theta}})$, that is, X_θ and $X_{\tilde{\theta}}$ do not have the same distribution. We now prove a stronger identifiability result that does not assume that $\tilde{\theta}$ is a restricted causal model.

Proposition 4.1 (Identifiability of causal additive tree models)**.** *Suppose that X_θ and $X_{\tilde{\theta}}$ are generated by the SCMs $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ and $\tilde{\theta} = (\tilde{\mathcal{G}}, (\tilde{f}_i), \tilde{P}_N) \in \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$, respectively. It holds that*

$$\mathcal{L}(X_\theta) = \mathcal{L}(X_{\tilde{\theta}}) \implies \mathcal{G} = \tilde{\mathcal{G}}.$$

We prove Proposition 4.1 using the techniques by Peters et al. (2014). While we prove the statement only for causal additive tree models, which suffices for this work, we conjecture that a similar extension holds for restricted structural causal DAG models. The extension of Proposition 4.1 is important for the following reason. Given a finite data set, practical methods usually assume that the true distribution is induced by an underlying restricted SCM. One can then fit different causal structures and output the structure that fits the data best. The above extension accounts for the fact that regression methods hardly represent all such restrictions: e.g., most nonlinear regression techniques can also fit linear models.

4.2.2. Score Functions

We now define population score functions which are later used to recover the causal tree. We henceforth assume that $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ is a random vector with distribution $P_X = X(P)$ generated by a causal additive tree model $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ with $\mathcal{G} = (V, \mathcal{E}) \in \mathcal{T}_p$ such that $\mathbb{E}\|X\|_2^2 < \infty$. Thus, $\mathcal{G}$ denotes the causal tree. We use $\tilde{\mathcal{G}} \in \mathcal{T}_p$ to denote an arbitrary, different (directed) tree. For the remainder of this paper, we assume that for any $i \neq j$ it holds that $X_i - \mathbb{E}[X_i|X_j]$ has a density with respect to Lebesgue measure.[3] We often refer to one of the following two scenarios: either, *(i)*, we have limited a priori information that $P_N \in \mathcal{P}_{+\mathcal{C}_3}^p$, or, *(ii)*, we know that the noise innovations are Gaussian, that is, $P_N \in \mathcal{P}_{\mathrm{G}}^p$. Whenever the data-generating noise distributions are Gaussian, we refer to this model as a Gaussian setup (or setting or model), even though the full distribution is not.

Definition 4.3. *For any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$ we define for each node $i \in V$ the*

[3]This ensures that the entropy score function introduced in Definition 4.3 below is well-defined and that the analysis of the identifiability gap in Section 4.5 is valid.

(i) local Gaussian score as $\ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) := \log\left(\mathrm{Var}\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right]\right)\right)/2$,

(ii) local entropy score as $\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) := h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right]\right)$,

(iii) local conditional entropy score as $\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) := h\left(X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right)$.

Here, we use the convention that $\mathbb{E}(X_i|\emptyset) = 0$ *and* $h(X_i|\emptyset) = h(X_i)$*; the functions* $h(\cdot)$*,* $h(\cdot|\cdot)$*, and* $h(\cdot,\cdot)$ *(used below) denote the differential entropy, conditional entropy, and cross entropy, respectively. The Gaussian, entropy and conditional entropy score of* $\tilde{\mathcal{G}}$ *are, respectively, given by the sum of local scores:*

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) := \sum_{i=1}^{p} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i), \quad \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) := \sum_{i=1}^{p} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i), \quad \ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) := \sum_{i=1}^{p} \ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i).$$

(See Polyanskiy and Wu (2019) or Cover and Thomas (2006) for the basic information-theoretic concepts used in this paper.)

The following lemma shows that the Gaussian score of the graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$ arises naturally as a translated infimum cross entropy between P_X and all Q induced by Gaussian SCMs. Similarly, the entropy score can be seen as an infimum cross entropy between P_X and all Q induced by another class of SCMs.

Lemma 4.2. *For any* $\tilde{\mathcal{G}} \in \mathcal{T}_p$ *it holds that*

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - p \log(\sqrt{2\pi e}).$$

Furthermore, with $\mathcal{F}(\tilde{\mathcal{G}}) := (\mathcal{F}_i(\tilde{\mathcal{G}}))_{1 \le i \le p}$*, where*

$$\mathcal{F}_i(\tilde{\mathcal{G}}) := \{x \mapsto \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)} = x]\},$$

for all $1 \le i \le p$*, it holds that*

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q).$$

Score-based methods identify the underlying structure by evaluating the score functions (or estimates thereof) on different graphs and choosing the best scoring graph. The difference between the score $\ell_\cdot(\mathcal{G})$ of the true graph and the score $\ell_\cdot(\tilde{\mathcal{G}})$ of the best scoring alternative graph $\tilde{\mathcal{G}}$ is an important property of the problem: e.g., if it would be zero, we could not identify the true graph from the scores. We, therefore, refer to expressions of the form $\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_\cdot(\tilde{\mathcal{G}}) - \ell_\cdot(\mathcal{G})$ as the identifiability gap.

In the remainder of this paper, we work under the assumption that the identifiability gap is strictly positive (see also Section 4.5).

Assumption 4.1. *If* $\theta \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ *or* $\theta \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ *it holds that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) > 0 \quad \textit{or} \quad \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) > 0, \tag{4.2}$$

respectively.

Assumption 4.1 does not trivially follow from the results further above. By arguments similar to those in Lemma 4.2 we have that, if the true data-generating model is a restricted Gaussian additive tree model, $\theta \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$, then $\ell_{\mathrm{G}}(\mathcal{G}) = h(P_X) - p\log(\sqrt{2\pi e})$. Hence, the Gaussian score gap between $\tilde{\mathcal{G}}$ and the causal graph $\mathcal{G}$ equals

$$\begin{aligned}\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - h(P_X) \\ &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} D_{\mathrm{KL}}(P_X \| Q),\end{aligned}$$

where D_{KL} denotes the Kullback-Leibler divergence measure. Proposition 4.1 implies that

$$\forall \tilde{\mathcal{G}} \neq \mathcal{G}, \quad \forall Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p : D_{\mathrm{KL}}(P_X \| Q) > 0.$$

However, this does not immediately imply that the identifiability gap (where we take the infimum over such Q) is strictly positive. Similar considerations [4] hold for the entropy score gap

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\mathrm{KL}}(P_X \| Q).$$

In Section 4.5 we derive informative lower bounds on the Gaussian and entropy score gaps (i.e., the infimum KL-divergence) of Equation (4.2). It is also possible to enforce Assumption 4.1 indirectly by the assumptions and modifications detailed in the following remark.

Remark 4.1. If $\theta \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$, such that for all $i \neq j$ it hold that $x \mapsto \mathbb{E}[X_i | X_j = x]$ has a differentiable version, then the Gaussian identifiability gap is strictly positive, so the first part of Assumption 4.1 holds. If $\theta \in \Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ and, in addition to the above condition it holds that for all $i \neq j$, $X_i - \mathbb{E}[X_i | X_j]$ has a continuous density, then then the entropy identifiability gap is strictly positive, so the in second part of Assumption 4.1 holds.

Assumption 4.1 can also be enforced by adopting the model restrictions of Bühlmann et al. (2014). Assume that $\Theta_R \subseteq \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ satisfies the further restriction that for all causal edges $(j \to i) \in \mathcal{E}$ the causal functions f_i are contained within a function class $\mathcal{F}_i \subseteq \mathcal{D}_1$ that is closed with respect to the $L^2(P_{X_j})$-norm. Now consider a modified Gaussian score function $\ell_{\mathrm{G.mod}} : \mathcal{T}_p \to \mathbb{R}$ that coincides with ℓ_{G} except that the conditional expectation function is replaced with $\arg\min_{f' \in \mathcal{F}_i} \mathbb{E}[(X_i - f'(X_j))^2] \in \mathcal{F}_i$. It now follows that

$$\ell_{\mathrm{G.mod}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G.mod}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times (\mathcal{F}_i)_{1 \leq i \leq p} \times \mathcal{P}_{\mathrm{G}}^p} D_{\mathrm{KL}}(P_X \| Q) > 0,$$

[4] In fact, Proposition 4.1 does not immediately imply that $D_{\mathrm{KL}}(P_X \| Q) > 0$ for $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ as it does not necessarily hold that the causal functions in $\mathcal{F}(\tilde{\mathcal{G}})$ are differentiable or that the noise innovation densities in $\mathcal{P}^p$ are continuous.

where the strict inequality follows from Proposition 4.1 as the infimum is attained for some $Q^* \in \{\tilde{\mathcal{G}}\} \times (\mathcal{F}_i)_{1\leq i\leq p} \times \mathcal{P}_{\mathrm{G}}^p$. Our theory and subsequent results transfer effortlessly to these modifications. ◦

We can now use the score functions to identify the true causal graph of a restricted structural model. In the Gaussian case, for example, we have, by virtue of Assumption 4.1,

$$\mathcal{G} = \operatorname*{arg\,min}_{\tilde{\mathcal{G}}\in\mathcal{T}_p} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}). \tag{4.3}$$

In practice, we consider estimates of the above quantities and optimize the corresponding empirical loss function. Solving Equation (4.3) (or its empirical counterpart) using exhaustive search is computationally intractable already for moderately large choices of p.[5] We now introduce CAT, a computationally efficient method that solves the optimization exactly.

4.3. Causal Additive Trees (CAT)

We introduce the population version of our algorithm CAT in Section 4.3.1 and discuss its finite sample version and asymptotic properties in Sections 4.3.2 and 4.3.3.

4.3.1. An Oracle Algorithm

Similarly as for the case of DAGs, the problem in Equation (4.3) is a combinatorial optimization problem, for which the cardinality of the search space grows super-exponentially with p. Indeed, the number of undirected trees on p labelled nodes is p^{p-2} (Cayley, 1889) and therefore p^{p-1} is the corresponding number of labelled trees. For the class of DAGs (which includes directed trees), existing structure learning such as Bühlmann et al. (2014) propose a greedy search technique that iteratively selects the lowest scoring directed edge under the constraint that no cycles is introduced in the resulting graph. In general, greedy search procedures do not come with any guarantees and there are indeed situations in which they fail (Peters et al., 2022). By exploiting the assumption of a tree structure, we will see that the optimization problem of Equation (4.3) can be solved computationally efficiently without the need for heuristic optimization techniques.

Provided with a connected directed graph with edge weights, Chu–Liu–Edmonds' algorithm finds a minimum edge weight directed spanning tree, given that such a tree exists. That is, for a connected directed graph $\mathcal{H} = (V, \mathcal{E}_{\mathcal{H}})$ on the nodes

[5] In the context of linear Gaussian models, Chickering (2002) proves consistency of greedy equivalent search towards the correct Markov equivalence class. This, however, does not imply that the optimization problem in Equation (4.3) is solved: for a given sample, the method is not guaranteed to find the optimal scoring graph (but the output will converge to the correct graph).

$V = \{1, \ldots, p\}$ with edge weights $\mathbf{w} := \{w(j \to i) : j \neq i\}$, Chu–Liu–Edmonds' algorithm recovers a minimum edge weight spanning directed tree subgraph of $\mathcal{H}$,

$$\underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p\cap\mathcal{H}}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w(j \to i),$$

where $\mathcal{T}_p \cap \mathcal{H}$ denotes all directed spanning trees of $\mathcal{H}$. The runtime of the original algorithms of Chu and Liu (1965) and Edmonds (1967) for a pre-specified root node is $\mathcal{O}(|\mathcal{E}_\mathcal{H}| \cdot p) = \mathcal{O}(p^3)$. Tarjan (1977) devised a modification of the algorithm that for dense graphs $\mathcal{H}$ and an unspecified root node has runtime $\mathcal{O}(p^2)$. In our experiments, we use the `C++` implementation of Tarjans modification by Tofigh and Sjölund (2007) which is contained in the R-package `RBGL` (Carey et al., 2021).

The causal graph recovery problem in Equation (4.3) is equivalently solved by finding a minimum edge weight directed tree, i.e., a minimum edge weight directed spanning tree of the fully connected graph on the nodes V. For example, finding the minimum of the Gaussian score function is equivalent to minimizing a translated version of the Gaussian score function

$$\begin{aligned}
&\underset{\tilde{\mathcal{G}}\in\mathcal{T}_p}{\arg\min}\, \ell_\mathrm{G}(\tilde{\mathcal{G}}) \\
&= \underset{\tilde{\mathcal{G}}\in\mathcal{T}_p}{\arg\min} \sum_{i=1}^{p} \frac{1}{2} \log(\mathrm{Var}(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}])) - \sum_{i=1}^{p} \frac{1}{2} \log(\mathrm{Var}(X_i)) \\
&= \underset{\tilde{\mathcal{G}}\in\mathcal{T}_p}{\arg\min} \sum_{i=1}^{p} \frac{1}{2} \log\left(\frac{\mathrm{Var}(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}])}{\mathrm{Var}(X_i)}\right).
\end{aligned}$$

Because the summand for the root note equals zero, we only need to sum over all nodes with an incoming edge in $\tilde{\mathcal{G}}$:

$$\mathcal{G} = \underset{\tilde{\mathcal{G}}\in\mathcal{T}_p}{\arg\min}\, \ell_\mathrm{G}(\tilde{\mathcal{G}}) = \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w_\mathrm{G}(j \to i),$$

for a Gaussian data-generating model. That is, the causal directed tree is given by the minimum edge weight directed tree with respect to the Gaussian edge weights $\mathbf{w}_\mathrm{G} := \{w_\mathrm{G}(j \to i) : j \neq i\}$ given by

$$w_\mathrm{G}(j \to i) := \frac{1}{2} \log\left(\frac{\mathrm{Var}(X_i - \mathbb{E}[X_i | X_j])}{\mathrm{Var}(X_i)}\right) \tag{4.1}$$

for all $j \neq i$. Similarly, the minimum of the entropy score function is given by the minimum edge weight directed tree with respect to the entropy edge weights $\mathbf{w}_\mathrm{E} := \{w_\mathrm{E}(j \to i) : j \neq i\}$ given by $w_\mathrm{E}(j \to i) := h(X_i - \mathbb{E}[X_i | X_j]) - h(X_i)$, for all $j \neq i$. We will henceforth denote the method where we apply Chu–Liu–Edmonds' algorithm on Gaussian and entropy edge weights as CAT.G and CAT.E, respectively.

4.3.2. Finite Sample Algorithm

Given an $n \times p$ data matrix $\mathbf{X}_n$, representing n i.i.d. copies of $X = (X_1, \ldots, X_p)$, we estimate the edge weights by simple plug-in estimators. Let us denote the conditional expectation function and its estimate by

$$\varphi_{ji}(x) := \mathbb{E}[X_i | X_j = x], \qquad \hat{\varphi}_{ji}(x) := \hat{\mathbb{E}}[X_i | X_j = x], \tag{4.2}$$

for any $j \neq i$. The estimated Gaussian edge weights are then given by

$$\hat{w}_{\mathrm{G}}(j \to i) := \frac{1}{2} \log \left(\frac{\widehat{\mathrm{Var}}(X_i - \hat{\varphi}_{ji}(X_j))}{\widehat{\mathrm{Var}}(X_i)} \right), \tag{4.3}$$

for all $i \neq j$, where $\widehat{\mathrm{Var}}(\cdot)$ denotes a variance estimator using the sample $\mathbf{X}_n$. We now propose to combine the Chu–Liu–Edmonds' algorithm described above with the Gaussian score as detailed in Algorithm 4.1.

Algorithm 4.1 Causal additive trees (CAT)

1: **procedure** CAT($\mathbf{X}_n$, regression method)
2: For each combination of (i, j) with $j \neq i$, run regression method to obtain $\hat{\varphi}_{ji}$.
3: Compute empirical edge weights $\hat{\mathbf{w}}_{\mathrm{G}} := (\hat{w}_{\mathrm{G}}(j \to i))_{j \neq i}$, see Equation (4.3).
4: Apply Chu–Liu–Edmonds' algorithm to the empirical edge weights.
5: **return** minimum edge weight directed tree $\hat{\mathcal{G}}$.
6: **end procedure**

By default we suggest to use the estimated Gaussian edge weights as described in Algorithm 4.1. However, it is also possible to run Chu–Liu–Edmonds' algorithm on estimated entropy edge weights given by

$$\hat{w}_{\mathrm{E}}(j \to i) := \hat{h}(X_i - \hat{\varphi}_{ji}(X_j)) - \hat{h}(X_i),$$

for all $j \neq i$, where $\hat{h}(\cdot)$ denotes a user-specific entropy estimator using the observed data $\mathbf{X}_n$. Estimating differential entropy is a difficult statistical problem but we will later in Section 4.6 demonstrate by simulation experiments that it can be beneficial to use the estimated entropy edge weights when the additive noise distributions are highly non-Gaussian.

Under suitable conditions on the (possibly nonparametric) regression technique, we now show that the proposed algorithm consistently recovers the true causal graph in Gaussian settings using estimated Gaussian edge weights.

4.3.3. Consistency

We study a version of the CAT.G algorithm applied to a Gaussian noise model where the regression estimates are trained on auxiliary data, simplifying the theoretical

analysis. We believe that consistency, with careful analysis, is achievable without sample splitting. As such, we only view the sample splitting as a theoretical device for simplifying proofs but we do not recommend it in practical applications. For each n we let $\mathbf{X}_n = (X_1, \ldots, X_n)$ and $\tilde{\mathbf{X}}_n = (\tilde{X}_1, \ldots, \tilde{X}_n)$ denote independent datasets each consisting of n i.i.d. copies of $X \in \mathbb{R}^p$. We suppose that the regression estimates $\hat{\varphi}_{ji}$ have been trained on $\tilde{\mathbf{X}}_n$ and then compute the edge weights using $\mathbf{X}_n$ as in step 3 of Algorithm 4.1:

$$\hat{w}_{\mathrm{G}}(j \to i) := \hat{w}_{ji}(\mathbf{X}_n, \tilde{\mathbf{X}}_n) := \frac{1}{2} \log \left(\frac{\frac{1}{n}\sum_{k=1}^{n} \left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j})\right)^2}{\frac{1}{n}\sum_{k=1}^{n} X_{k,i}^2 - (\frac{1}{n}\sum_{k=1}^{n} X_{k,i})^2} \right). \tag{4.4}$$

The following result shows pointwise consistency of CAT.G whenever the conditional mean estimation is weakly consistent.

Theorem 4.1 (Pointwise consistency)**.** *Suppose that for all $j \neq i$ the following two conditions hold:*

(a) if $(j \to i) \in \mathcal{E}$, $\mathbb{E}[(\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$;

(b) if $(j \to i) \notin \mathcal{E}$, $\mathbb{E}[(\hat{\varphi}_{ji}(X_j) - \tilde{\varphi}_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$ for some fixed $\tilde{\varphi}_{ji} : \mathbb{R} \to \mathbb{R}$.

Here, φ_{ji} and $\hat{\varphi}_{ji}$ are defined in Equation (4.2)*. In the large sample limit, we recover the causal graph with probability one, that is*

$$P(\hat{\mathcal{G}} = \mathcal{G}) \to_n 1,$$

where $\hat{w}_{\mathrm{G}}(j \to i)$ is given by Equation (4.4)*.*

The assumptions of Theorem 4.1 only require weakly consistent estimation of the conditional means for edges that are present in the causal graph; these represent causal relationships and are often assumed to be smooth. This distinction allow us to employ regression techniques that are consistent only for those function classes that we consider reasonable for modeling the causal mechanisms. For non-causal edges, $(j \to i) \notin \mathcal{E}$, the estimator $\hat{\varphi}_{ji}$ only needs to converge to a function $\tilde{\varphi}_{ji}$, which does not necessarily need to be the conditional mean.

4.3.3.1. Consistency under Vanishing Identifiability

We now consider an asymptotic regime involving a sequence $(\theta_n)_{n \in \mathbb{N}}$ of a sequence of SCMs with potentially changing conditional mean functions φ_{ji} and a vanishing identifiability gap. We have the following result.

Theorem 4.2 (Consistency under vanishing identifiability)**.** *Let $(\theta_n)_{n \in \mathbb{N}}$ be a sequence of SCMs on $p \in \mathbb{N}$ nodes all with the same causal directed tree $\mathcal{G} = (V, \mathcal{E})$ such that*

(i) for $q_n := \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\mathcal{G}) - \ell_{\mathrm{G}}(\tilde{\mathcal{G}})$ (the gap of model θ_n), we have $q_n^{-1} = o(\sqrt{n})$;

(ii) for all $(j \to i) \in \mathcal{E}$ *and* $\varepsilon > 0$,

$$P_{\theta_n}\left(q_n^{-1}\mathbb{E}_{\theta_n}\left[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^2|\tilde{\mathbf{X}}_n\right] > \varepsilon\right) \to_n 0;$$

(iii) for all $j \neq i$ *and* $\varepsilon > 0$,

$$P_{\theta_n}\left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^4|\tilde{\mathbf{X}}_n\right] > \varepsilon\right) \to_n 0; \text{ and}$$

(iv) there exists $C > 0$ *such that for all* $j \neq i$

$$\inf_n P_{\theta_n}(\mathrm{Var}_{\theta_n}(X_i|X_j) \leq C) = 1 \text{ and } \sup_n \mathbb{E}_{\theta_n}\|X\|_2^4 < \infty.$$

Then it holds that

$$P(\hat{\mathcal{G}} = \mathcal{G}) \to_n 1.$$

Condition (i) asks that the identifiability gap q_n goes to zero more slowly than the standard convergence rate $1/\sqrt{n}$ of estimators in regular parametric models. Such a requirement would be necessary in almost any structure identification problem. Condition (ii) requires the mean squared error of the regression estimates corresponding to true causal edges to be $o_P(q_n)$. We regard this as a fairly mild assumption: indeed, the minimax rate of estimation of regression functions in Hölder balls with smoothness β is $n^{-2\beta/(2\beta+1)}$ (Tsybakov, 2009). Thus, we can expect that if the causal regression functions have smoothness $\beta \geq 1/2$ and all lie in a Hölder ball, (ii) can be satisfied for any q_n satisfying (i). Condition (iii) allows the fourth moments of the estimation errors to increase at any rate slower than $nq_n^2 \to \infty$; of course, we would typically expect this error to decay, at least for the causal edges.

4.4. Hypothesis Testing

This section presents a procedure to test any substructure hypothesis regarding the causal directed tree of a Gaussian additive noise model. We continue our analysis using the sample split estimators of Equation (4.4), where the conditional expectations are estimated on an auxiliary dataset. Our approach makes use of the fact that the estimated weights in Equation (4.4) are logarithms of ratios of i.i.d. quantities, and thus the joint distribution of the estimated edge weights should, with appropriate centering and scaling, be asymptotically Gaussian; see Lemma C.4 in Appendix C.4 for the precise statement. This allows us to create a (biased) confidence region of the true edge weights, which in turn gives a confidence set for the true graph. This confidence set of graphs is not necessarily straightforward to compute and list. However, we show that it can be queried to test hypotheses of interest, such as the presence or absence of a particular edge. As these hypothesis

tests are derived from a confidence region, they are valid even when the hypothesis to test has been chosen after examining the data.

Similar to the results in the previous sections, we avoid making assumptions on the performance of regressions corresponding to non-causal edges. Unlike the consistency analysis, however, here we do not require identifiability of the true graph, but in the non-identified case all assumptions and conclusions below involving the 'true graph' should be interpreted as involving the set of all population score minimizing graphs.

In order to state our results we introduce the following notation. For a collection $(K_{ji})_{j\neq i}$, we let $K_i := (K_{1i}, \dots, K_{(i-1)i}, K_{(i+1)i}, \dots, K_{pi})^\intercal \in \mathbb{R}^{p-1}$, furthermore, for any collection $(K_i)_{1\leq i\leq p}$, we let $K := (K_1, \dots, K_p)^\intercal$. With this notation, let the vectors of squared residuals and squared centered observations be given by

$$\hat{M}_k := \{(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\}_{j\neq i} \in \mathbb{R}^{p(p-1)},$$
$$\hat{V}_k := \left\{\left(X_{i,k} - \frac{1}{n}\sum_{k=1}^{n} X_{i,k}\right)^2\right\}_{1\leq i\leq p} \in \mathbb{R}^p.$$

Further let

$$\hat{\mu} := \frac{1}{n}\sum_{k=1}^{n} \hat{M}_k, \qquad \hat{\nu} =: \frac{1}{n}\sum_{k=1}^{n} \hat{V}_k.$$

Note that with this notation, the estimated Gaussian edge weight for $j \to i$ is given by $\log(\hat{\mu}_{ji}/\hat{\nu}_i)/2$. Let us denote by $\widehat{\Sigma}_M \in \mathbb{R}^{p(p-1)\cdot p(p-1)}$, $\widehat{\Sigma}_V \in \mathbb{R}^{p\cdot p}$ and $\widehat{\Sigma}_{MV} \in \mathbb{R}^{p(p-1)\cdot p}$, the empirical variances of the $\hat{M}_k$ and $\hat{V}_k$ and their empirical covariance respectively, so

$$\begin{pmatrix} \widehat{\Sigma}_M & \widehat{\Sigma}_{MV} \\ \widehat{\Sigma}_{MV}^\intercal & \widehat{\Sigma}_V \end{pmatrix} := \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} \hat{M}_k\hat{M}_k^\intercal - \hat{\mu}\hat{\mu}^\intercal & \hat{M}_k\hat{V}_k^\intercal - \hat{\mu}\hat{\nu}^\intercal \\ \hat{V}_k\hat{M}_k^\intercal - \hat{\nu}\hat{\mu}^\intercal & V_kV_k^\intercal - \hat{\nu}\hat{\nu}^\intercal \end{pmatrix}.$$

With this, we may now present our construction of confidence intervals for the edge weights. (For simplicity, all proofs in this section assume the variables to have mean zero.)

4.4.1. Confidence Region for the Causal Tree

We use the delta method to estimate the variances of the $\hat{w}_{ji}$, and a simple Bonferroni correction to ensure simultaneous coverage of the confidence intervals we develop. Writing z_α for the upper $\alpha/\{2p(p-1)\}$ quantile of a standard normal distribution, we set

$$\hat{u}_{ji},\, \hat{l}_{ji} := \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right) \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}},$$

where

$$\hat{\sigma}_{ji}^2 := \frac{\widehat{\Sigma}_{M,ji,ji}}{\hat{\mu}_{ji}^2} + \frac{\widehat{\Sigma}_{V,i,i}}{\hat{\nu}_i^2} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{\hat{\mu}_{ji}\hat{\nu}_i}.$$

We treat $[\hat{l}_{ji}, \hat{u}_{ji}]$ as a confidence interval for the true edge weight $w_{\mathrm{G}}(j \to i)$ and define the following region of directed trees formed of minimizers of the score with edge weights in the confidence hyperrectangle:

$$\hat{C} := \left\{ \operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji}, : \forall j \neq i, w'_{ji} \in [\hat{l}_{ji}, \hat{u}_{ji}] \right\}.$$

We have the following coverage guarantee for $\hat{C}$.

Theorem 4.3 (Confidence region)**.** *Suppose the following conditions hold:*

(i) there exists $\xi > 0$ such that $\mathbb{E}\|X\|^{4+\xi} < \infty$;

(ii) there exists $\xi > 0$ such that for all $j \neq i$, $\mathbb{E}[|\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j)|^{4+\xi}|\tilde{\mathbf{X}}_n] = O_p(1)$;

(iii) $\mathrm{Var}((\hat{M}_1^\intercal, \hat{V}_1^\intercal)^\intercal|\tilde{\mathbf{X}}_n) \xrightarrow{P}_n \Sigma$, where Σ is constant with strictly positive diagonal;

(iv) for $(j \to i) \in \mathcal{E}$, $\sqrt{n}\mathbb{E}[(\hat{\varphi}_{ji}(X_{k,j}) - \varphi_{ji}(X_{k,j}))^2|\tilde{\mathbf{X}}_n] \xrightarrow{P} 0$.

Then

$$\liminf_{n\to\infty} P\left(\mathcal{G} \in \hat{C}\right) \geq 1 - \alpha.$$

The second condition requires little more than 4th moments for the absolute errors in the regression (they do not need to converge to zero). Condition (iv) requires that the mean squared prediction errors corresponding to the true causal edges decay faster than a relatively slow $1/\sqrt{n}$ rate. If the causal graph is unidentifiable, then when (iv) holds for all edges corresponding to population score minimizing graphs, $\hat{C}$ will cover every such graph with a probability of at least $1 - \alpha$.

4.4.2. Testing of Substructures

Whilst the confidence region $\hat{C}$ has attractive coverage properties, it will typically not be possible to compute it in practice. We now introduce a computationally feasible scheme for querying whether $\hat{C}$ satisfies certain constraints such as containing or not containing a given substructure. A substructure $\mathcal{R} = (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\mathrm{miss}}, r)$ on the nodes V contains specified sets $\mathcal{E}_{\mathcal{R}}$ and $\mathcal{E}_{\mathcal{R}}^{\mathrm{miss}}$ of existing and missing edges, respectively, and/or a specific root node r; for example, this could be a specific directed tree or a single edge (such as $X_1 \to X_2$) or a single missing edge (such as $X_1 \not\to X_2$). Our approach allows us to report with certainty that at least one of the constraints in $\mathcal{R}$ does *not* hold for the true graph. More precisely, we propose a test for the null hypothesis

$$\mathcal{H}_0(\mathcal{R}) : \mathcal{E}_{\mathcal{R}} \setminus \mathcal{E} = \emptyset,\ \mathcal{E} \setminus \mathcal{E}_{\mathcal{R}}^{\mathrm{miss}} = \emptyset,\ r = \mathrm{rt}(\mathcal{G}),$$

i.e, that all constraints in $\mathcal{R}$ are satisfied in the causal graph.

In order to present our method, we introduce some notation. Let $s(w)$ be the score attained by the minimum edge weight directed tree recovered by Chu–Liu–Edmonds' algorithm with input edge weights $w := (w_{ji})_{j \neq i}$. Let $\mathcal{T}(\mathcal{R}) \subseteq \mathcal{T}_p$ be the set of all directed trees satisfying the constraints $\mathcal{R}$. Furthermore, let $s_{\mathcal{T}(\mathcal{R})}(w)$ be the score attained by the minimum edge weight directed tree in $\mathcal{T}(\mathcal{R})$. Now suppose that the causal directed tree $\mathcal{G}$ satisfies the constraints $\mathcal{R}$. If $[\hat{l}, \hat{u}] := \prod_{j \neq i} [\hat{l}_{ji}, \hat{u}_{ji}]$ is an asymptotically valid confidence region for the Gaussian population edge weights $\mathbf{w}_{\mathrm{G}}$ defined in Equation (4.1), we have with probability tending to $1 - \alpha$ that

$$s_{\mathcal{T}(\mathcal{R})}(\hat{l}) \leq s_{\mathcal{T}(\mathcal{R})}(\mathbf{w}_{\mathrm{G}}) = s(\mathbf{w}_{\mathrm{G}}) \leq s(\hat{u}).$$

We may thus set as our test function

$$\psi_{\mathcal{R}} = \mathbb{1}_{\{s_{\mathcal{T}(\mathcal{R})}(\hat{l}) > s(\hat{u})\}}.$$

The expressions $s_{\mathcal{T}(\mathcal{R})}(\hat{l})$ and $s(\hat{u})$ can be computed from the data. For $s_{\mathcal{T}(\mathcal{R})}(\hat{l})$, we perform the following steps: we apply Chu–Liu–Edmonds' algorithm on the edge weights $\hat{l}$ where, for any $(j \to i) \in \mathcal{E}_{\mathcal{R}}$, we remove all other edges into i from the edge pool (or set the corresponding edge weight to sufficiently large values) while for a specified root node $r \in \mathcal{R}$ we remove all incoming edges into r from the edge pool. Edges $(j \to i) \in \mathcal{E}_{\mathcal{R}}^{\mathrm{miss}}$ are removed from the edge pool, too.

Formalizing a line of reasoning similar to the above, taking into account that $[\hat{l}, \hat{u}]$ is in fact a biased confidence region that may not necessarily contain the population edge weights with increasing probability, we have the following result.

Theorem 4.4 (Pointwise asymptotic level)**.** *Suppose that the conditions of Theorem 4.3 are satisfied and let $\mathcal{R}_1, \mathcal{R}_2, \ldots$ be any collection of potentially data-dependent constraints. For any level $\alpha \in (0, 1)$, we have that*

$$\limsup_{n \to \infty} P\left(\bigcup_{k \,:\, \mathcal{H}_0(\mathcal{R}_k) \text{ is true}} \{\psi_{\mathcal{R}_k} = 1\} \right) \leq \alpha.$$

4.5. Bounding the Identifiability Gap

We have seen that the identifiability gap, that is, the smallest score difference between the causal tree $\mathcal{G}$ and any alternative graph $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, plays an important role when identifying causal trees from data. It provides information about whether the causal graph is identifiable by means of the corresponding score function, and it affects how quickly the estimation error needs to vanish in order to guarantee consistency, see Theorem 4.2. E.g., for the entropy score, the

identifiability gap is given by

$$\begin{aligned}\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \sum_{i=1}^{p} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i) \\ &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\mathrm{KL}}(P_X \| Q), \end{aligned} \tag{4.1}$$

see Section 4.2.2.

We now analyze the identifiability gap for the entropy score and the Gaussian score in more detail. More specifically, we will derive a lower bound for the identifiability gaps that is based on local properties of the underlying structural causal models (such as the ability to reverse edges). We first consider the special cases of bivariate models (Section 4.5.1) and multivariate Markov equivalent trees (Section 4.5.2) and then turn to general trees (Section 4.5.3). However, before we venture into the derivation of the specific lower bounds we first examine the connection between the identifiability gaps associated with the different score functions.

In this section, we assume that $X \sim P_X$ is generated by a structural causal additive tree model with $\mathbb{E}\|X\|^2 < \infty$ such that the local Gaussian, entropy and conditional entropy scores are well-defined. We neither assume that θ is a restricted structural causal additive model, i.e., $\theta \in \Theta_R$, nor strict positivity of the identifiability gap, i.e., Assumption 4.1. The following result shows that the local node-wise score gaps associated with the different score functions are ordered.

Lemma 4.3. *For any $\tilde{\mathcal{G}} \in \mathcal{T}_p$ and for all $i \in V$*

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{CE}}(\mathcal{G}, i) \leq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i). \tag{4.2}$$

If the underlying model is a Gaussian noise model, then

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i) \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{G}}(\mathcal{G}, i). \tag{4.3}$$

It follows that the full graph score gaps and identifiability gaps associated with the different score functions satisfy a similar ordering. Thus, given that the underlying model is Gaussian, a strictly positive entropy identifiability gap implies that the Gaussian identifiability gap is strictly positive. It is, however, not possible to establish strict positivity of the conditional entropy identifiability gap; see Remark C.1 in Appendix C.2. Therefore, we focus on establishing a lower bound for the entropy identifiability gap that is tighter than that given by the conditional entropy identifiability gap.

In general, we cannot use node-wise comparisons of the scores of two graphs to bound the identifiability gap (the reason is that in general a node receives a better score in a graph, where it has a parent, compared to a graph, where it does not; see Example C.1 in Appendix C.2 for a formal argument). We start by analyzing the identifiability gap in models with two variables.

4.5.1. Bivariate Models

We now consider two nodes $V = \{X, Y\}$, and graphs $\mathcal{T}_2 = \{(X \to Y), (Y \to X)\}$. Without loss of generality assume that $(X, Y) \in \mathcal{L}^2(P)$ is generated by an additive noise SCM $\theta = (\mathcal{G}, (f_i), P_N)$ with causal graph $\mathcal{G} = (X \to Y) \in \mathcal{T}_2$ to which the only alternative graph is $\tilde{\mathcal{G}} = (Y \to X)$. That is,

$$X := N_X, \quad Y := f(X) + N_Y, \tag{4.4}$$

where $(N_X, N_Y) \sim P_N \in \mathcal{P}^2$. The bivariate entropy identifiability gap, which we will later refer to as the edge reversal entropy score gap, is defined as

$$\begin{aligned}\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) :&= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \\ &= h(Y) + h(X - \mathbb{E}[X|Y]) - h(X) - h(Y - \mathbb{E}[Y|X]),\end{aligned}$$

where the fully drawn arrow symbolizes the true causal relationship and the dashed arrow the alternative. The following lemma simplifies the bivariate entropy identifiability gap to a single mutual information between the effect and the residual of the minimum mean squared prediction error regression of cause on the effect.

Lemma 4.4. *Consider the bivariate setup of Equation* (4.4) *and assume that* $f(X)$ *has density. It holds that*

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) = I(X - \mathbb{E}[X|Y]; Y) \geq 0.$$

Thus, the causal graph is identified in a bivariate setting if one maintains dependence between the predictor and minimum mean squared error regression residual in the anti-causal direction. This result is in accordance with the previous identifiability results. For example, in the linear Gaussian case, $I(X - \mathbb{E}[X|Y]; Y) = 0$. Consequently, the causal graph is not identified from the entropy score function.

Whenever the conditional mean in the anti-causal direction vanishes, e.g., with symmetric causal function and symmetric noise distribution, it is possible to derive a more explicit lower bound with more intuitive sufficient conditions for identifiability of the causal graph.

Proposition 4.2. *Consider the bivariate setup of Equation* (4.4) *and assume that* $f(X)$ *has density. If the reversed direction conditional mean* $\mathbb{E}[X|Y]$ *almost surely vanishes (e.g., because* f*,* X *and* N_Y *are symmetric), then*

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) = I(X; f(X) + N_Y),$$

which is strictly positive if and only if $X \not\perp\!\!\!\perp f(X) + N_Y$*. In addition, we have the following statements.*

(a) Let $f(X)^{\mathrm{G}}$ *and* N_Y^{G} *be independently normal distributed with the same mean and variance as* $f(X)$ *and* N_Y*, respectively. If*

$$D_{\mathrm{KL}}(f(X) \| f(X)^{\mathrm{G}}) \leq D_{\mathrm{KL}}(N_Y \| N_Y^{\mathrm{G}}),$$

then

$$\Delta\ell_{\mathrm{E}}(X \xleftarrow{} Y) \geq \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).$$

(b) If the density of $f(X) + N_Y$ is log-concave, then

$$\Delta\ell_{\mathrm{E}}(X \xleftarrow{} Y) \geq \frac{1}{2}\log\left(\frac{2}{\pi e} + \frac{2}{\pi e}\frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).$$

This lower bound is non-trivial only if

$$\mathrm{Var}(f(X)) > (\pi e/2 - 1)\mathrm{Var}(N_Y) \approx 3.27\mathrm{Var}(N_Y).$$

Thus, if the conditional mean $\mathbb{E}[X|Y]$ in the anti-causal direction vanishes, then under certain conditions, the causal direction is identified by the entropy score function (as long as $\mathrm{Var}(f(X))$ is sufficiently large relative to $\mathrm{Var}(N_Y)$). The edge reversal score gap for the Gaussian score is given by

$$\begin{aligned}\Delta\ell_{\mathrm{G}}(X \xleftarrow{} Y) &:= \frac{1}{2}\log\left(\frac{\mathrm{Var}(X - \mathbb{E}[X|Y])}{\mathrm{Var}(X)}\right) - \frac{1}{2}\log\left(\frac{\mathrm{Var}(Y - \mathbb{E}[Y|X])}{\mathrm{Var}(Y)}\right) \\ &= \frac{1}{2}\log\left(\frac{\mathrm{Var}(X - \mathbb{E}[X|Y])}{\mathrm{Var}(X)}\right) + \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right),\end{aligned}$$

which reduces to the lower bound in point (a) of Proposition 4.2 if the conditional mean $\mathbb{E}[X|Y]$ in the anti-causal direction vanishes.

4.5.2. Multivariate Markov Equivalent Trees

Two Markov equivalent trees differ in precisely one directed path that is reversed in one graph relative to the other.[6] The entropy score gap of Markov equivalent trees therefore reduces to the binary case.

Proposition 4.3. *Consider any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$ that is Markov equivalent to the causal tree $\mathcal{G}$. Let $c_1 \to \cdots \to c_r$ be the unique directed path in $\mathcal{G}$ that is reversed in $\tilde{\mathcal{G}}$. Then*

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \sum_{i=1}^{r-1} \Delta\ell_{\mathrm{E}}(c_i \xleftarrow{} c_{i+1}) \geq \min_{1\leq i\leq r-1} \Delta\ell_{\mathrm{E}}(c_i \xleftarrow{} c_{i+1}).$$

Thus, a lower bound of the entropy score gap that holds uniformly over the Markov equivalence class is given by the smallest possible edge reversal in the causal directed graph:

$$\min_{\tilde{\mathcal{G}}\in\mathrm{MEC}(\mathcal{G})\setminus\{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(j\to i)\in\mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{} i).$$

[6] To see this, note that any two directed trees are Markov equivalent if and only if they satisfy the exact same d-separations or equivalently they share the same skeleton (there are no v-structures in directed trees). Distinct directed trees sharing the same skeleton must have distinct root nodes. Consequently, there exist a directed path in $\mathcal{G}$ from $\mathrm{rt}(\mathcal{G})$ to $\mathrm{rt}(\tilde{\mathcal{G}})$ that is reversed in $\tilde{\mathcal{G}}$; see also Lemma C.6

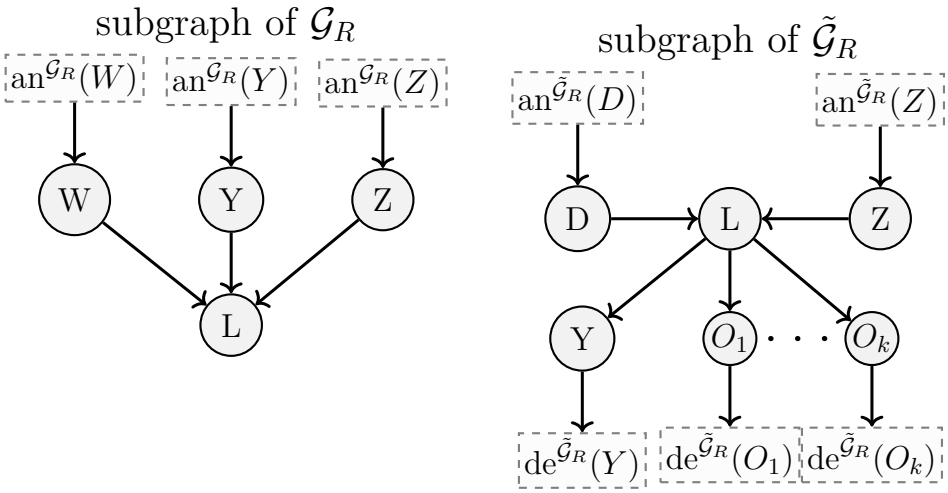


Figure 4.1: Schematic illustration of parts of two reduced graphs produced by the graph reduction technique described in Section 4.5.3. Consider a sink node L in $\mathcal{G}_R$. Its parent (in $\mathcal{G}_R$) must either be a parent in $\tilde{\mathcal{G}}_R$, too, it must be a child in $\tilde{\mathcal{G}}_R$, or it is unconnected to L in $\tilde{\mathcal{G}}_R$. Thus, exactly one of the sets Z, Y, and W is non-empty. This case distinction is used to compute the three bounds in Theorem 4.5. D, $O_1, \ldots, O_k$ denote further (possibly existing) nodes in $\tilde{\mathcal{G}}_R$.

4.5.3. General Multivariate Trees

We now derive a lower bound of the entropy identifiability gap, i.e., a lower bound of the entropy score gap that holds uniformly over all alternative trees $\mathcal{T}_p \setminus \{\mathcal{G}\}$. To do so, we exploit a graph reduction technique (introduced by Peters et al., 2014) which enables us to reduce the analysis to three distinct scenarios. This graph reduction works as follows. Fix any alternative graph $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, and iteratively remove any node (from both $\mathcal{G}$ and $\tilde{\mathcal{G}}$) that has no children and the same parents in both $\mathcal{G}$ and $\tilde{\mathcal{G}}$. The score gap is unaffected by the graph reduction.[7]

Applying this iteration scheme, until no such node can be found, results in two reduced graphs $\mathcal{G}_R = (V_R, \mathcal{E}_R)$ and $\tilde{\mathcal{G}}_R = (V_R, \tilde{\mathcal{E}}_R)$. These reduced graphs cannot be empty, for that would only happen if $\tilde{\mathcal{G}} = \mathcal{G}$. Further, they have identical vertices but different edges. And they can be categorized into one of three cases. To do so, consider a node L that is a sink node, i.e., a node without children, in $\mathcal{G}_R$ and consider its parent in $\mathcal{G}_R$. Now, considering $\tilde{\mathcal{G}}_R$, one of the following conditions must hold: the parent is also a parent of L in $\tilde{\mathcal{G}}_R$ (we then call it Z), the parent is not connected to L in $\tilde{\mathcal{G}}_R$ (we then call it W), or the parent is a child of L in $\tilde{\mathcal{G}}_R$ (we then call it Y). Figure 4.1 visualizes these three scenarios.

[7] All removed nodes $V \setminus V_R$ have identical incoming edges in both graphs and therefore have identical local scores. That is, for any loss function $l \in \{\ell_{\text{CE}}, \ell_{\text{E}}, \ell_{\text{G}}\}$ we have that $l(\tilde{\mathcal{G}}) - \ell(\mathcal{G}) = \sum_{i \in V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) + \sum_{i \in V \setminus V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) = \sum_{i \in V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) = \ell(\tilde{\mathcal{G}}_R) - l(\mathcal{G}_R)$.

We can now obtain bounds for each of the three case individually. For the case with a node Z (a 'staying parent'), define

$$\Pi_Z(\mathcal{G}) := \left\{(z,l,o) \in V^3 \text{ s.t. } (z \to l) \in \mathcal{E} \text{ and } o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z,l\}\right\}.$$

The score gap can then be lower bounded by $\min_{(z,l,o)\in\Pi_Z(\mathcal{G})} I(X_z; X_o|X_l)$ (see Lemma C.7). Intuitively, $I(X_z; X_o|X_l)$ quantifies the strength of the connection between z and o, when conditioning on l (which does not lie on the path between z and o). This is a non-local bound in that it does not constrain the length of the path connecting z and o. Analyzing or bounding this term might be difficult. We will see in 4.5.4 that this part is not needed in the Gaussian case.

For the case with a node W ('removing parent'), define

$$\Pi_W(\mathcal{G}) := \left\{(w,l,o) \in V^3 \text{ s.t. } (w \to l) \in \mathcal{E}, o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)\right\}.$$

This case results in the lower bound $\min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l|X_o)$ (see Lemma C.8). Here, w is a parent of l and o is directly connected to w. Intuitively, $I(X_w; X_l|X_o)$ quantifies the strength of the edge $w \to l$. We condition on o but that node is not directly connected to l (only via w). For the first two cases, faithfulness (Spirtes et al., 2000) implies that these terms are non-zero and bounding them away from zero reminds of strong faithfulness (Zhang and Spirtes, 2002). However, in the second case, one considers individual edges, which reminds more of a strong version of causal minimality (Peters et al., 2017; Spirtes et al., 2000).

For the case with a node Y ('parent to child'), a lower bound is given by the minimal edge reversal score gap $\min_{(j\to i)\in\mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{-\!-\!>} i)$ (see Lemma C.9). The term $\Delta\ell_{\mathrm{E}}(j \xleftarrow{-\!-\!>} i)$ measures the identifiability of the direction of an individual edge. It is zero in the linear Gaussian case, for example. We provide more details on the reduced graphs and on the arguments in the three cases in Section C.4.4.2 of Appendix C.4.

Combining the three bounds from above, we obtain the following theorem.

Theorem 4.5. *It holds that*

$$\min_{\tilde{\mathcal{G}}\in\mathcal{T}_p\setminus\{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min\Bigg\{ \min_{(z,l,o)\in\Pi_Z(\mathcal{G})} I(X_z; X_o|X_l), \min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l|X_o), \min_{(j\to i)\in\mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{-\!-\!>} i)\Bigg\}. \tag{4.5}$$

This result lower bounds the identifiability gap using information-theoretic quantities. Corresponding results for the Gaussian score follow immediately by Lemma 4.3. The last two terms are local properties of the underlying structural causal model; the first term is not. As seen in Section 4.5.2, the last term on the right-hand side is required when considering only Markov equivalent trees; if it is non-zero, it allows us to orient all edges in the skeleton. The first two terms

(non-zero under faithfulness) are additionally required when the considered trees are not Markov equivalent.

We now turn to the case of Gaussian trees. Here, the first term is not needed; the bound then depends only on local properties of the structural causal model.

4.5.4. Gaussian Multivariate Trees

The score gap lower bound in Equation (4.5) consists of local dependence properties except for the node tuples $\Pi_Z(\mathcal{G})$ (Lemma C.7) that arise when considering alternative graphs that yield in reduced graphs with a node Z ('staying parents'). However, we show that in the Gaussian case, the score gap for such alternative graphs can be lower bounded by the score gaps already considered in alternative graphs with a node Y ('parent to child') and a node W ('removing parent'). Thus, we have the following theorem, with a bound consisting only of local properties of the model.

Theorem 4.6 (Gaussian localization of the identifiability gap)**.** *In a Gaussian setting (see Section 4.2.2), we have*

$$\begin{aligned}&\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) \\ &\geq \min\left\{ \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o), \min_{(j \to i) \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{\cdot\!\cdot\!\cdot} i) \right\}.\end{aligned}$$

4.6. Simulation Experiments

In this section, we investigate the finite-sample performance of CAT and perform simulation experiments investigating the identifiability gap and its lower bound. In Section 4.6.1 we compare the performance of CAT to CAM of Bühlmann et al. (2014) for Gaussian and non-Gaussian additive noise models with causal graphs given by directed trees. In Section 4.6.2 we perform simulation experiments that highlight the behavior of the identifiability gap and its corresponding lower bound derived in Section 4.5. In Section 4.6.3 we compare the CAT and CAM for causal discovery on non-tree DAG models (CAT always outputs a directed tree). The code scripts (R) for the simulation experiments and an implementation of CAT is available on GitHub.[8]

4.6.1. Causal Structure Learning for Trees

In this section, we compare the performance of the structure learning methods CAT and CAM when employed on additive noise models with causal graphs given by directed trees.

[8] `https://github.com/MartinEmilJakobsen/CAT`

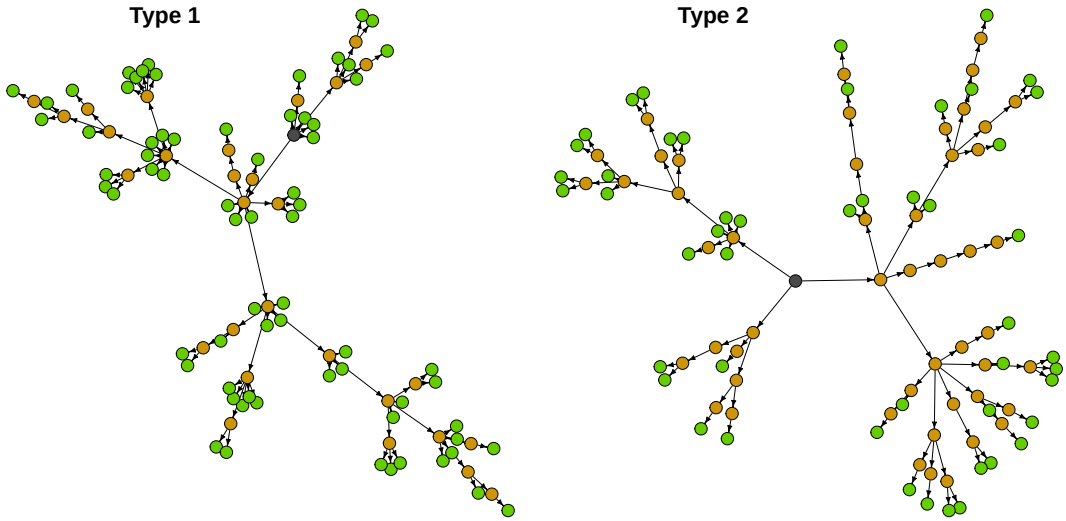


Figure 4.2: Illustration of Type 1 (many leaf nodes) and Type 2 (many branch nodes) directed trees over $p = 100$ nodes. The green nodes are leaf nodes, the brown nodes are branch nodes, and the black nodes are root nodes. The Type 1 tree contains 70 leaf nodes, while the Type 2 tree only contains 49 leaf nodes.

4.6.1.1. Tree Generation Schemes

We employ two different random directed tree generation schemes: Type 1 (many leaf nodes) and Type 2 (many branch nodes). In Figure 4.2 we have illustrated two directed trees generated in accordance with the two generation schemes. For more details, see Algorithms C.1 and C.2 in Section C.3.1 of Appendix C.3.

4.6.1.2. Gaussian Experiment

In this experiment, we generate data similarly to the experimental setup of Bühlmann et al. (2014). For any given directed tree we generate causal functions by sample paths of Gaussian processes with radial basis function (RBF) kernel and bandwidth parameter of one. Sample paths of Gaussian processes with radial basis function kernels are almost surely infinitely continuous differentiable (e.g., Kanagawa et al., 2018), non-constant and nonlinear, so they satisfy the requirements of Lemma 4.1. See Figure C.1 in Section C.3.2 of Appendix C.3 for illustrations of random draws of such functions. Root nodes are mean zero Gaussian variables with standard deviation sampled uniformly on $(1, 2)$. Furthermore, for each fixed tree and set of causal functions, we introduce at each non-root node additive Gaussian noise with mean zero and standard deviation sampled uniformly on $(1/5, \sqrt{2}/5)$.

We first compare our method CAT with Gaussian score function (CAT.G) against the method CAM of Bühlmann et al. (2014) on the previously detailed

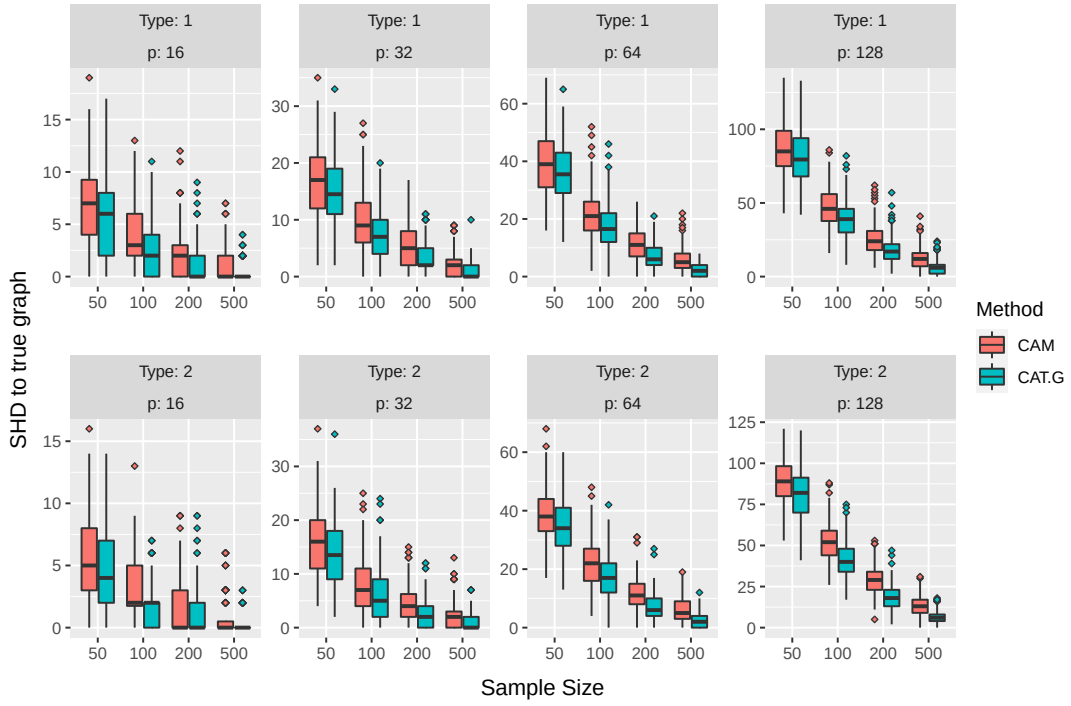


Figure 4.3: Gaussian setting: Boxplots of the SHD performance of CAM and CAT.G (Gaussian score) for varying sample sizes, system sizes, and tree types. CAT.G outperforms CAM in a wide range of scenarios.

nonlinear additive Gaussian noise tree setup. We implement CAT.G without sample-splitting and use the R-package `GAM` (Generalized Additive Models, Hastie, 2020) with default settings to construct a thin plate regression spline estimate of the conditional expectations. We use the implementation of Chu–Liu–Edmonds' algorithm from the R-package `RBGL`.[9] CAM is employed with a maximum number of parents set to one (restricting the output to directed trees), without preliminary neighborhood selection and subsequent pruning. We measure the performance of the methods by computing the Structural Hamming Distance (SHD, Tsamardinos et al., 2006) and Structural Intervention Distance (SID, Peters and Bühlmann, 2015) to the causal tree.

For each system size $p \in \{16, 32, 64, 128\}$ we generate a causal tree, corresponding causal functions and noise variances and sample data of size $n \in \{50, 100, 200, 500\}$. This is repeated 200 times and the SHD results are summarized in the boxplot of Figure 4.3. Both methods perform better on trees of Type 2 than on trees of Type 1. CAT.G outperforms CAM in terms of SHD to the true graph both in

[9]The RBGL implementation finds maximum edge weight directed trees and requires all positive edge weights. As such, we take the negative of our edge weights and shift them all by the absolute value of smallest edge-weight. If an edge weight is set to zero this edge can not be chosen.

median distance and IQR length and position for all sample sizes, system sizes and tree types. Considering the SID to the causal tree yields similar conclusions; see Figure C.2 in Section C.3.2 of Appendix C.3. In their default versions, CAM and CAT.G use different estimation techniques of the conditional expectations, but this does not seem to be the source of the performance difference: Figure C.3 in Section C.3.2 of Appendix C.3 illustrates a similar SHD performance difference when forcing CAT.G to use the edge weights produced by the CAM implementation.

4.6.1.3. Non-Gaussian Experiment

We now compare the performance of CAM and CAT with Gaussian (CAT.G) and entropy (CAT.E) score functions in a setup with varying noise distributions. The entropy edge weights used by CAT.E are estimated with the differential entropy estimator of Berrett et al. (2019) as implemented in the CRAN R-package `IndepTest` (Berrett et al., 2018). We use the same simulation setup as in Section 4.6.1.2 but now we only consider trees of Type 1 and parameterize the setup by $\alpha > 0$, which controls the deviation of the additive noise innovations from a Gaussian distribution. More precisely, we generate the additive noise variables $N_i(\alpha)$ as

$$N_i(\alpha) = \text{sign}(Z_i)|Z_i|^\alpha,$$

where $Z_i \sim \mathcal{N}(0, \sigma_i^2)$ with σ_i sampled uniformly on $(1/5, \sqrt{2}/5)$ or uniformly on $(1, 2)$ if $i = \text{rt}(\mathcal{G})$. For $\alpha = 1$ this yields Gaussian noise, while for alpha $\alpha \neq 1$ the noise is non-Gaussian. We conduct the experiment for all combinations of $\alpha \in \{0.1, 0.2, \ldots, 2, 2.5, 3, 3.5, 4\}$ and sample sizes $n \in \{50, 500\}$ for a fixed system size of $p = 32$. Each setting is repeated 500 times and the results are illustrated in Figure 4.4.

For Gaussian noise, both CAM and CAT.G outperform CAT.E. This can (at least) be attributed to two factors: (i) CAT.E does not, unlike CAM and CAT.G, explicitly use the Gaussian noise specification and (ii) differential entropy estimation is a difficult statistical problem (see, e.g., Han et al., 2020; Paninski, 2003) For small and moderate deviations from Gaussianity, CAT.G outperforms both CAM and CAT.E. For larger deviations, CAT.E outperforms both CAT.G and CAM in terms of median SHD. Finally, we note that CAT.G always outperforms CAM in terms of median SHD.

4.6.2. Identifiability Gap

We now investigate the behavior of the identifiability gap in bivariate models (Section 4.6.2.1) and evalute the lower bound derived in Section 4.5 empirically for multivariate models (Section 4.6.2.2).

4.6.2.1. Bivariate Identifiability Gap

In this experiment, we investigate the behavior of the bivariate identifiability gap and analyze both a Gaussian and a non-Gaussian setup. Let us consider an additive

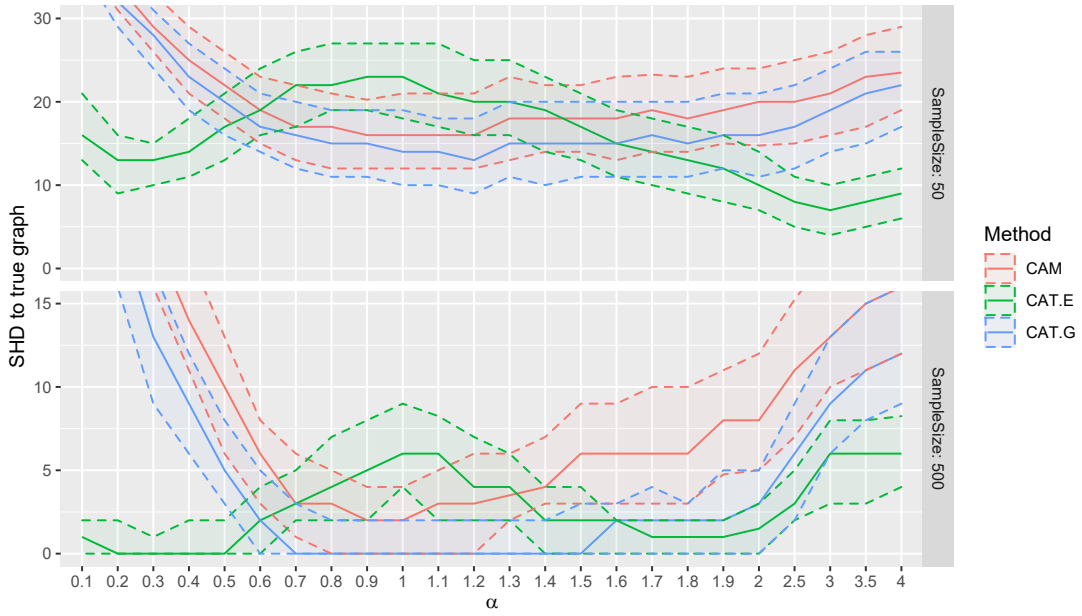


Figure 4.4: Deviations from Gaussianity: The parameter α controls the noise deviation from the Gaussian distribution. CAT.G and CAT.E are instances of CAT with edge weights derived from Gaussian and entropy score functions, respectively. The solid lines represent the median SHD and the shaded (dashed) region represents the interquartile range. Using the entropy score yields better results for noise distributions that deviate strongly from Gaussian noise.

noise model over (X, Y) with causal graph $X \to Y$. The causal functions will be chosen from the following function class. For any $\lambda \in [0, 1]$, define $f_\lambda : \mathbb{R} \to \mathbb{R}$ as

$$f_\lambda(x) = (1 - \lambda)x^3 + \lambda x.$$

That is, $\lambda \mapsto f_\lambda$ interpolates between a cubic function $x \mapsto x^3$ and a linear function $x \mapsto x$. For any $(\alpha, \lambda) \in (0, \infty) \times [0, 1]$ we consider the following bivariate structural causal additive model

$$X := \operatorname{sign}(N_X)|N_X|^\alpha, \quad Y := f_\lambda(X) + N_Y,$$

where N_X, N_Y are independent standard normal distributed random variables. Recall that the bivariate identifiability gap is given by

$$\begin{aligned} \ell_{\mathrm{E}}(Y \to X) - \ell_{\mathrm{E}}(X \to Y) &= h(X - \mathbb{E}[X|Y]) + h(Y) - h(X - \mathbb{E}[X|Y], Y) \\ &= I(X - \mathbb{E}[X|Y]; Y), \end{aligned} \tag{4.1}$$

by Lemma 4.4. Thus, the causal graph $X \to Y$ is identified by the entropy score function if $I(X - \mathbb{E}[X|Y]; Y) > 0$.

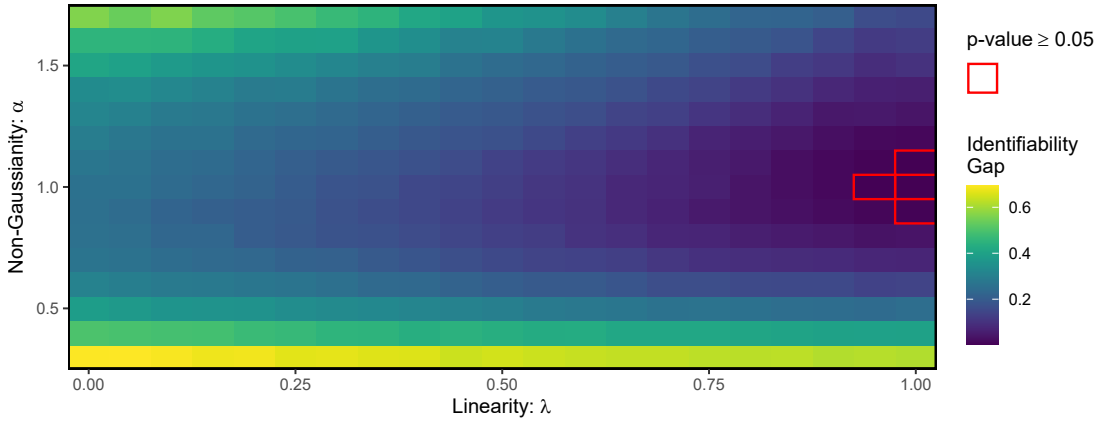


Figure 4.5: Heatmap of the identifiability gap for varying λ and α. Tiles with a red boundary correspond to the models for which the mutual information based independence test cannot reject the null hypothesis of a vanishing identifiability gap.

For any fixed λ and α we now estimate the identifiability gap; we also calculate the p-value associated with the null hypothesis that the identifiability gap is zero (based on 50000 observations). Similarly to the previous experiment, we estimate the conditional expectations using GAM. We estimate (without sample splitting) the identifiability gap and construct p-values using the CRAN R-package `IndepTest` (Berrett et al., 2018). More specifically, we use the differential entropy estimator of Berrett et al. (2019) and the mutual information based independence test of Berrett and Samworth (2019), respectively.

The heatmap of Figure 4.5 illustrates the behavior of the identifiability gap for all combinations of $\lambda \in \{0, 0.05, \ldots, 1\}$ and $\alpha \in \{0.3, 0.4, \ldots, 1.7\}$. It suggests that the identifiability gap only tends to zero when we approach the linear Gaussian setup. Only in the models closest to the linear Gaussian setup are we unable to reject the null-hypothesis of a vanishing identifiability gap.

This is also what the theory predicts, namely that for bivariate linear Gaussian additive models, the causal direction is not identified. It is known that for linear models, non-Gaussianity is helpful for identifiability. The empirical results indicate that the same holds for nonlinear models, i.e., that the identifiability gap increases with the degree of non-Gaussianity.

4.6.2.2. Multivariate Identifiability Gap

In this experiment, we investigate the identifiability gap and its relation to the lower bounds established in Theorem 4.6. For a Gaussian additive noise tree model,

it holds that

$$\begin{aligned}&\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) \\ &\geq \min \left\{ \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o), \min_{i \to j \in \mathcal{E}} \Delta \ell_{\mathrm{E}}(i \xleftarrow{} j) \right\}.\end{aligned}$$

In other words, the identifiability gap is lower bounded by the minimum of the smallest local faithfulness measures and the smallest edge-reversal score difference. We now investigate empirically how important the first term is for the inequality to hold. More specifically, for a given model generation scheme, we quantify how often the minimum edge reversal is sufficiently small to establish the lower bound without the conditional mutual information term, that is, how often the identifiability constant $\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G})$ is larger than the minimum edge reversal.

The minimum edge reversal can be estimated using the same conditional expectation and entropy estimators of the experiment in Section 4.6.2.1. However, estimating the identifiability gap between the second-best scoring tree and the causal tree needs further elaboration. We know that the best scoring (causal) tree can be found by Chu–Liu–Edmonds' (a directed MWST) algorithm. The second-best scoring tree differs from the best scoring tree in at least one edge. Thus, given the best scoring graph, we remove one of the $p-1$ edges of the best scoring tree from the pool of possible edges and rerun Chu–Liu–Edmonds' algorithm. We do this for each of the $p-1$ edges in the best scoring tree which leaves us with $p-1$ possibly different sub-optimal trees of which the minimum score is attained by the second-best scoring graph.

For the experiment, we randomly sample data generating models similarly to the experiment in Section 4.6.1.2. However, we change the causal functions from explicit sample paths of a Gaussian process to a GAM model estimating the sample paths due to memory constraints when generating large sample sizes. Figure 4.6 illustrates, for $p \in \{8, 16\}$, boxplots of the difference between the identifiability gap and the minimum edge reversal for 100 randomly generated Gaussian additive noise tree models. For each model, the identifiability gap and corresponding minimum edge reversal is estimated from 200000 independent and identically distributed observations. The illustration suggests that it is in general necessary to also consider the conditional mutual information term in order to establish a lower bound. However, it also shows that in the majority (90%) of the models, the minimum edge reversal is indeed a lower bound for the identifiability gap.

4.6.3. Robustness: CAT on DAGs

This experiment analyzes how CAT performs compared to CAM when applied to data generated from a Gaussian additive model with a non-tree DAG as a causal graph. More specifically, we analyze the behavior on single-rooted DAGs. For any fixed $p \in \mathbb{N}$ we generate a directed tree of Type 1 and for

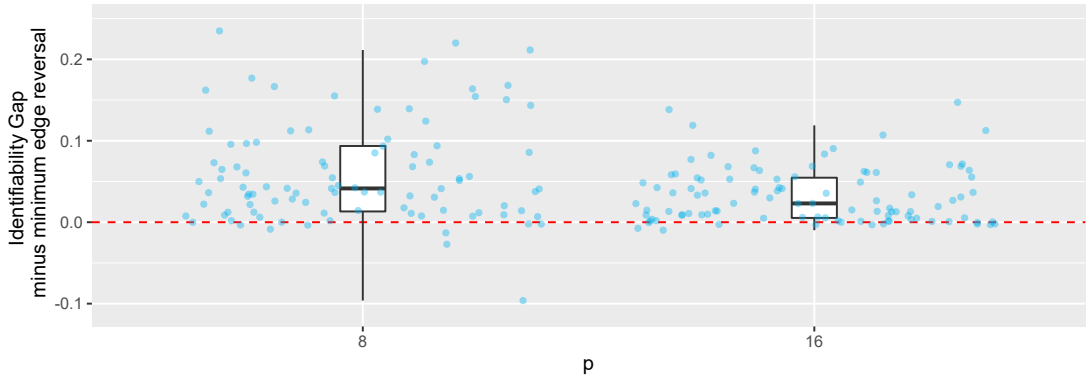


Figure 4.6: Empirical analysis of the lower bound on the identifiability gap, see Section 4.6.2.2. In most of the simulated settings, we see that the estimated identifiability gap is larger than the smallest edge-reversal score difference. This suggests that in many cases, the latter term is sufficient for establishing a lower bound on the identifiability gap.

each zero in the upper triangular part of the adjacency matrix we add an edge with 5% probability. The causal functions and Gaussian noise innovations are generated according to the specifications given in the experiment of Section 4.6.2.2. The structural assignment for each node is additive in each causal parent, i.e., for all $i \in \{1, \dots, p\}$, $X_i := \sum_{j \in \mathrm{pa}^{\mathcal{G}}(i)} f_{ji}(X_j) + N_i$, with $(N_1, \dots, N_p)$ mutually independent Gaussian distributed noise innovations. For each $p \in \{16, 32, 64\}$ and sample size $n \in \{50, 250, 500\}$ we randomly generate 100 single-rooted Gaussian additive models according to the above specifications.

As CAT.G outputs trees, we do not expect it to output the correct graph. Figure 4.7 illustrates the performance of CAT.G and CAM in terms of ancestor relations. For this experiment, we employ CAM with preliminary neighborhood selection and subsequent pruning. For small systems, CAM slightly outperforms CAT.G in terms of true positive rate (TPR) when classifying causal ancestors. However, for large systems and large sample sizes, CAT.G outperforms CAM in that metric. On the other hand, CAM is not limited to trees which allows it to find a more significant proportion of the true ancestor, as seen by the fraction of correctly classified ancestors over actual ancestors. CAT.G seems to be a viable alternative for practical non-tree applications where the true positive rate of estimated ancestors is more important than finding all ancestor relations.

In Figure C.4 of Section C.3.2 of Appendix C.3 we have illustrated similar comparisons when focusing on recovered edges. The true positive rate of the recovered edges for CAT.G is larger than CAM only for small sample sizes, while the opposite is true for large sample sizes. As expected, and as for the ancestor relationships, the fraction of correctly predicted edges over total causal edges is significantly higher for CAM.

Finally, while both methods are relatively efficient, CAT has a slightly lower

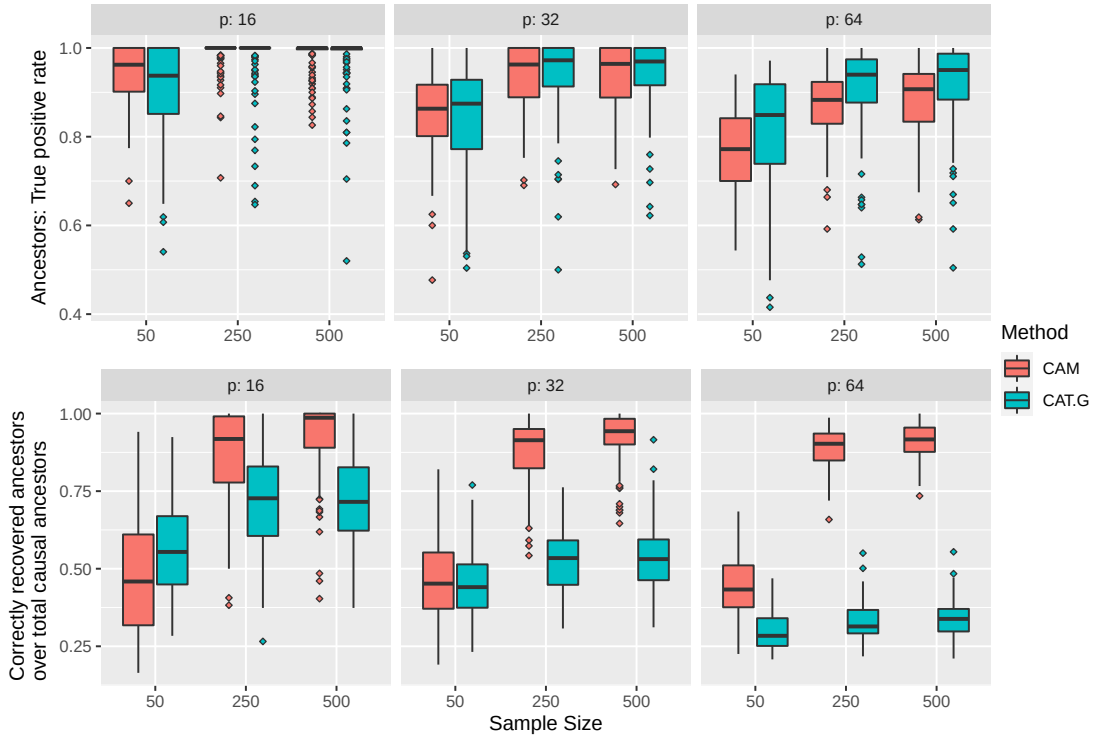


Figure 4.7: Estimating ancestor relations in non-tree DAGs, see Section 4.6.3. CAT.G slightly outperforms CAM in terms of true positive rates for large graphs (top) but finds less ancestor relationships (bottom) due to fitting a tree.

runtime than the greedy search algorithm of CAM. The average runtime of CAM and CAT.G in this experiment for $p = 64$ and $n = 500$ was 193 and 139 seconds, respectively. For both methods, the most time consuming part is estimating of the conditional expectations that are used to compute the edge weights.

4.7. Summary and Future Work

This paper shows that exact structure learning is possible for systems of lesser complexity, i.e., for restricted structural causal models with additive noise and causal graphs given by directed trees. We propose the method CAT, which is guaranteed to consistently recover the causal directed tree in a Gaussian noise setting under mild assumptions on the regression methods used to estimate conditional means. Furthermore, we argue that CAT is consistent in an asymptotic setup with vanishing identifiability. We present a computationally feasible procedure to test substructure hypotheses and provide an analysis of the identifiability gap. Simulation experiments show that CAT outperforms other (more general) structure learning methods for the specific task of recovering the causal graph in additive noise structural causal models when the causal structure is given by directed trees.

The proof of Proposition 4.1 is based on the fact that the causal functions of alternative models are differentiable and that the noise densities are continuous. We conjecture that it is possible to get even stronger identifiability statements under weaker assumptions; proving such a result necessitates new proof strategies. Furthermore, it should be possible to bootstrap a unbiased simultaneous hypercube confidence region for the Gaussian edge weights. This, however, requires a sufficiently fast convergence rate of the estimation error of the conditional expectations corresponding to non-causal edges. Compared to the Bonferroni correction, this approach could increase the power of the test.

Acknowledgments

We thank Phillip Bredahl Mogensen and Thomas Berrett for helpful discussions on the entropy score and its estimation. PB and JP thank David Bürge and Jan Ernest for helpful discussions on exploiting Chu–Liu–Edmonds' algorithm for causal discovery during the early stages of this project. MEJ and JP were supported by the Carlsberg Foundation; JP was, in addition, supported by a research grant (18968) from VILLUM FONDEN. RDS was supported by EPSRC grant EP/N031938/1. PB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461).

Chapter 5

Learning Summary Graphs of Time Series and Artifacts in DAG Models

Joint work with

Sebastian Weichwald, Phillip Bredahl Mogensen, Lasse Petersen, Nikolaj Thams and Gherardo Varando

Abstract

In this article, we describe the algorithms for causal structure learning from time series data that won the Causality 4 Climate competition at the Conference on Neural Information Processing Systems 2019 (NeurIPS). We examine how our combination of established ideas achieves competitive performance on semi-realistic and realistic time series data exhibiting common challenges in real-world Earth sciences data. In particular, we discuss a) a rationale for leveraging linear methods to identify causal links in non-linear systems, b) a simulation-backed explanation as to why large regression coefficients may predict causal links better in practice than small p-values and thus why normalising the data may sometimes hinder causal structure learning.

For benchmark usage, we detail the algorithms here and provide implementations at github.com/sweichwald/tidybench. We propose the presented competition-proven methods for baseline benchmark comparisons to guide the development of novel algorithms for structure learning from time series.

Keywords: Causal discovery, structure learning, time series, scaling.

5.1. Introduction

Inferring causal relationships from large-scale observational studies is an essential aspect of modern climate science Runge et al., 2019a,b. However, randomised studies and controlled interventions cannot be carried out, due to both ethical and practical reasons. Instead, simulation studies based on climate models are state-of-the-art to study the complex patterns present in Earth climate systems (IPCC, 2013).

Causal inference methodology can integrate and validate current climate models and can be used to probe cause-effect relationships between observed variables.

The Causality 4 Climate (C4C) NeurIPS competition (Runge et al., 2020) aimed to further the understanding and development of methods for structure learning from time series data exhibiting common challenges in and properties of realistic weather and climate data.

Structure of this work Section 5.2 introduces the structure learning task considered. In Section 5.3, we describe our winning algorithms. With a combination of established ideas, our algorithms achieved competitive performance on semi-realistic data across all 34 challenges in the C4C competition track. Furthermore, at the time of writing, our algorithms lead the rankings for all hybrid and realistic data set categories available on the CauseMe.net benchmark platform which also offers additional synthetic data categories (Runge et al., 2019b). These algorithms—which can be implemented in a few lines of code—are built on simple methods, are computationally efficient, and exhibit solid performance across a variety of different data sets. We therefore encourage the use of these algorithms as baseline benchmarks and guidance of future algorithmic and methodological developments for structure learning from time series.

Beyond the description of our algorithms, we aim at providing intuition that can explain the phenomena we have observed throughout solving the competition task. First, if we *only* ask whether a causal link exists in some non-linear time series system, then we may sidestep the extra complexity of explicit non-linear model extensions (cf. Section 5.4). Second, when data has a meaningful natural scale, it may—somewhat unexpectedly—be advisable to forego data normalisation and to use raw (vector auto)-regression coefficients instead of p-values to assess whether a causal link exists or not (cf. Section 5.5).

5.2. Causal Structure Learning from Time-discrete Observations

The task of inferring the causal structure from observational data is often referred to as 'causal discovery' and was pioneered by Pearl (2009) and Spirtes et al. (2000). Much of the causal inference literature is concerned with structure learning from independent and identically distributed (iid) observations. Here, we briefly review some aspects and common assumptions for causally modelling time-evolving systems. More detailed and comprehensive information can be found in the provided references.

Time-discrete observations We may view the discrete-time observations as arising from an underlying continuous-time causal system (Peters et al., 2020). While difficult to conceptualise, the correspondence between structural causal models and differential equation models can be made formally precise (Bongers and Mooij, 2018; Mooij et al., 2013; Rubenstein et al., 2018). Taken together, this

yields some justification for modelling dynamical systems by discrete-time causal models.

Summary graph as inferential target It is common to assume a time-homogeneous causal structure such that the dynamics of the observation vector X are governed by $X^t := F(X^{\text{past}(t)}, N^t)$ where the function F determines the next observation based on past values $X^{\text{past}(t)}$ and the noise innovation N^t. Here, structure learning amounts to identifying the summary graph with adjacency matrix A that summarises the causal structure in the following sense: the $(i, j)^{\text{th}}$ entry of the matrix A is 1 if $X_i^{\text{past}(t)}$ enters the structural equation of X_i^t via the i^{th} component of F and 0 otherwise. If $A_{ij} = 1$, we say that "X_i causes X_j". While summary graphs can capture the existence and non-existence of cause-effect relationships, they do in general not correspond to a time-agnostic structural causal model that admits a causal semantics consistent with the underlying time-resolved structural causal model (Janzing et al., 2018; Rubenstein et al., 2017).

Time structure may be helpful for discovery In contrast to the iid setting, the Markov equivalence class of the summary graph induced by the structural equations of a dynamical system is a singleton when assuming causal sufficiency and no instantaneous effects (Mogensen and Hansen, 2020; Peters et al., 2017). This essentially yields a justification and a constraint-based causal inference perspective on Wiener-Granger-causality (Granger, 1969; Peters et al., 2017; Wiener, 1956)

Challenges for causal structure learning from time series data Structure learning from time series is a challenging task hurdled by further problems such as time-aggregation, time-delays, and time-subsampling. All these challenges were considered in the C4C competition and are topics of active research (Danks and Plis, 2013; Hyttinen et al., 2016).

5.3. The Time-series Discovery Benchmark (tidybench): Winning Algorithms

We developed four simple algorithms,

- `SLARAC` Subsampled Linear Auto-Regression Absolute Coefficients (cf. Alg. 1)
- `QRBS` Quantiles of Ridge regressed Bootstrap Samples (cf. Alg. 2)
- `LASAR` LASso Auto-Regression
- `SELVAR` Selective auto-regressive model

which came in first in 18 and close second in 13 out of the 34 C4C competition categories and won the overall competition (Runge et al., 2020). Here, we provide

detailed descriptions of the `SLARAC` and `QRBS` algorithms. DYAnalogous descriptions for the latter two algorithms and implementations of all four algorithms are available at github.com/sweichwald/tidybench.

All of our algorithms output an edge score matrix that contains for each variable pair (X_i, X_j) a score that reflects how likely it is that the edge $X_i \to X_j$ exists. Higher scores correspond to edges that are inferred to be more likely to exist than edges with lower scores, based on the observed data. That is, we rank edges relative to one another but do not perform hypothesis tests for the existence of individual edges. A binary decision can be obtained by choosing a cut-off value for the obtained edge scores. In the C4C competition, submissions were compared to the ground-truth cause-effect adjacency matrix and assessed based on the achieved ROC-AUC when predicting which causal links exist.

The idea behind our algorithms is the following: regress present on past values and inspect the regression coefficients to decide whether one variable is a Granger-cause of another. `SLARAC` fits a VAR model on bootstrap samples of the data each time choosing a random number of lags to include; `QRBS` considers bootstrap samples of the data and Ridge-regresses time-deltas $X(t)-X(t-1)$ on the preceding values $X(t-1)$; `LASAR` considers bootstrap samples of the data and iteratively—up to a maximum lag—LASSO-regresses the residuals of the preceding step onto values one step further in the past and keeps track of the variable selection at each lag to fit an OLS regression in the end with only the selected variables at selected lags included; and `SELVAR` selects edges employing a hill-climbing procedure based on the leave-one-out residual sum of squares and finally scores the selected edges with the absolute values of the regression coefficients. In the absence of instantaneous effects and hidden confounders, Granger-causes are equivalent to a variable's causal parents (Peters et al., 2017, Theorem 10.3). In Section 5.5, we argue that the size of the regression coefficients may in certain scenarios be more informative about the existence of a causal link than standard test statistics for the hypothesis of a coefficient being zero. It is argued that for additive noise models, information about the causal ordering may be contained in the raw marginal variances. In test statistics such as the F- and T-statistics, this information is lost when normalising by the marginal variances.

5.4. Capturing Nonlinear Cause-Effect Links by Linear Methods

We explain the rationale behind our graph reconstruction algorithms and how they may capture non-linear dynamics despite being based on linearly regressing present on past values. For simplicity we will outline the idea in a multivariate regression setting with additive noise, but it extends to the time series setting by assuming time homogeneity.

Algorithm 1: Subsampled Linear Auto-Regression Absolute Coefficients (`SLARAC`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.
Parameters: Max number of lags, $L \in \mathbb{N}$.
Number of bootstrap samples, $B \in \mathbb{N}$.
Individual bootstrap sample sizes, $\{v_1, \ldots, v_B\}$.
Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

Initialise A_{full} as a $d \times dL$ matrix of zeros and $\widehat{A}$ as an empty $d \times d$ matrix;
for $b = 1, \ldots, B$ **do**
  lags $\leftarrow$ random integer in $\{1, \ldots, L\}$;
  Draw a bootstrap sample $\{t_1, \ldots, t_{v_b}\}$ from $\{\text{lags}+1, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1), \ldots \boldsymbol{X}(t_{v_b}))$;
  $\boldsymbol{X}^{(b)}_{\text{past}} \leftarrow \begin{pmatrix} \boldsymbol{X}(t_1 - 1) & \cdots & \boldsymbol{X}(t_1 - \text{lags}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}(t_{v_b} - 1) & \cdots & \boldsymbol{X}(t_{v_b} - \text{lags}) \end{pmatrix}$;
  Fit OLS estimate β of regressing $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}_{\text{past}}$;
  Zero-pad β such that $\dim \beta = d \times dL$;
  $A_{\text{full}} \leftarrow A_{\text{full}} + |\beta|$;
end
Aggregate $(\widehat{A})_{i,j} \leftarrow \max((A_{\text{full}})_{i,j+0\cdot d}, \ldots, (A_{\text{full}})_{i,j+L\cdot d})$ for every i, j;
Return: Score matrix $\widehat{A}$.

Let $N, X(t_1), X(t_2) \in \mathbb{R}^d$ be random variables such that

$$X(t_2) := F(X(t_1)) + N$$

for some differentiable function $F = (F_1, \ldots, F_d) : \mathbb{R}^d \to \mathbb{R}^d$. Assume that N has mean zero, that it is independent from $X(t_1)$, and that it has mutually independent components. For each $i, j = 1, \ldots, d$ we define the quantity of interest

$$\theta_{ij} = \mathbb{E}\left|\partial_i F_j\left(X(t_1)\right)\right|,$$

such that θ_{ij} measures the expected effect from $X_i(t_1)$ to $X_j(t_2)$. We take the matrix $\Theta = \left(\mathbf{1}_{\theta_{ij}>0}\right)$ as the adjacency matrix of the summary graph between $X(t_1)$ and $X(t_2)$.

In order to detect regions with non-zero gradients of F we create bootstrap samples $\mathcal{D}_1, \ldots, \mathcal{D}_B$. On each bootstrap sample $\mathcal{D}_b$ we obtain the regression coefficients $\widehat{A}_b$ as estimate of the directional derivatives by a (possibly penalised) linear regression technique. Intuitively, if θ_{ij} were zero, then on any bootstrap sample we would obtain a small non-zero contribution. Conversely, if θ_{ij} were

Algorithm 2: Quantiles of Ridge regressed Bootstrap Samples (`QRBS`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.

Parameters : Number of bootstrap samples, $B \in \mathbb{N}$.
Size of bootstrap samples, $v \in \mathbb{N}$.
Ridge regression penalty, $\kappa \geq 0$.
Quantile for aggregating scores, $q \in [0, 1]$.

Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

for $b = 1, \ldots, B$ **do**
  Draw a bootstrap sample $\{t_1, \ldots, t_v\}$ from $\{2, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1) - \boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v) - \boldsymbol{X}(t_v - 1))$;
  $\boldsymbol{X}^{(b)} \leftarrow (\boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v - 1))$;
  Fit a ridge regression of $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}$:
  $\widehat{A}_b = \arg\min_A \|\boldsymbol{Y}^{(b)} - A\boldsymbol{X}^{(b)}\| + \kappa\|A\|$;
Aggregate $\widehat{A} \leftarrow q^{th}$ element-wise quantile of $\{|\widehat{A}_1|, \ldots, |\widehat{A}_B|\}$;
Return Score matrix $\widehat{A}$.

non-zero, then we may for some bootstrap samples obtain a linear fit of $X_j(t_2)$ with large absolute regression coefficient for $X_i(t_1)$. The values obtained on each bootstrap sample are then aggregated by, for example, taking the average of the absolute regression coefficients $\widehat{\theta}_{ij} = \frac{1}{B}\sum_{b=1}^{B} \left|(\widehat{A}_b)_{ij}\right|$.

This amounts to searching the predictor space for an effect from $X_i(t_1)$ to $X_j(t_2)$, which is approximated linearly. It is important to aggregate the absolute values of the coefficients to avoid cancellation of positive and negative coefficients. The score $\widehat{\theta}_{ij}$ as such contains no information about whether the effect from $X_i(t_1)$ to $X_j(t_2)$ is positive or negative and it cannot be used to predict $X_j(t_2)$ from $X_i(t_1)$. It serves as a score for the existence of a link between the two variables. This rationale explains how linear methods may be employed for edge detection in non-linear settings without requiring extensions of Granger-type methods that explicitly model the non-linear dynamics and hence come with additional sample complexity (Marinazzo et al., 2008, 2011; Stramaglia et al., 2012, 2014).

5.5. Large Regression Coefficients May Predict Causal Links Better in Practice Than Small P-values

This section aims at providing intuition behind two phenomena: We observed a considerable drop in the accuracy of our edge predictions whenever 1) we normalised the data or 2) used the T-statistics corresponding to testing the hypothesis of

regression coefficients being zero to score edges instead of the coefficients' absolute magnitude. While one could try to attribute these phenomena to some undesired artefact in the competition setup, it is instructive to instead try to understand when exactly one would expect such behaviour.

We illustrate a possible explanation behind these phenomena and do so in an iid setting in favour of a clear exposition, while the intuition extends to settings of time series observations and our proposed algorithms. The key remark is, that under comparable noise variances, the variables' marginal variances tend to increase along the causal ordering. If data are observed at comparable scales—say sea level pressure in different locations measured in the same units—or at scales that are in some sense naturally relative to the true data generating mechanism, then absolute regression coefficients may be preferable to T-test statistics. Effect variables tend to have larger marginal variance than their causal ancestors. This helpful signal in the data is diminished by normalising the data or the rescaling when computing the T-statistics corresponding to testing the regression coefficients for being zero. This rationale is closely linked to the identifiability of Gaussian structural equation models under equal error variances Peters and Bühlmann (2014). Without any prior knowledge about what physical quantities the variables correspond to and their natural scales, normalisation remains a reasonable first step. We are not advocating that one should use the raw coefficients and not normalise data, but these are two possible alterations of existing structure learning procedures that may or may not, depending on the concrete application at hand, be worthwhile exploring. Our algorithms do not perform data normalisation, so the choice is up to the user whether to feed normalised or raw data, and one could easily change to using p-values or T-statistics instead of raw coefficients for edge scoring.

5.5.1. Instructive IID Case Simulation Illustrates Scaling Effects

We consider data simulated from a standard acyclic linear Gaussian model. Let $N \sim \mathcal{N}\left(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\right)$ be a d-dimensional random variable and let $\boldsymbol{B}$ be a $d \times d$ strictly lower-triangular matrix. Further, let X be a d-valued random variable constructed according to the structural equation $X = \boldsymbol{B}X + N$, which induces a distribution over X via $X = (I - \boldsymbol{B})^{-1}N$. We have assumed, without loss of generality, that the causal order is aligned such that X_i is further up in the causal order than X_j whenever $i < j$. We ran 100 repetitions of the experiment, each time sampling a random lower triangular 50×50-matrix $\boldsymbol{B}$ where each entry in the lower triangle is drawn from a standard Gaussian with probability $1/4$ and set to zero otherwise. For each such obtained $\boldsymbol{B}$ we sample $n = 200$ observations from $X = \boldsymbol{B}X + N$ which we arrange in a data matrix $\boldsymbol{X} \in \mathbb{R}^{200 \times 50}$ of zero-centred columns denoted by $\boldsymbol{X}_j$.

We regress each X_j onto all remaining variables $X_{\neg j}$ and compare scoring edges $X_i \to X_j$ by the absolute values of a) the regression coefficients $|\widehat{b}_{i \to j}|$, versus b) the T-statistics $|\widehat{t}_{i \to j}|$ corresponding to testing the hypothesis that the regression

coefficient $\widehat{b}_{i\to j}$ is zero. That is, we consider

$$|\widehat{b}_{i\to j}| = \left|(\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_{\neg j})^{-1}\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_j\right|_i$$

versus

$$|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}|\sqrt{\frac{\widehat{\mathrm{var}}(X_i|X_{\neg i})}{\widehat{\mathrm{var}}(X_j|X_{\neg j})}}\sqrt{\frac{(n-d)}{\left(1-\widehat{\mathrm{corr}}^2(X_i,X_j|X_{\neg\{i,j\}})\right)}} \tag{5.1}$$

where $\widehat{\mathrm{var}}(X_j|X_{\neg j})$ is the residual variance after regressing X_j onto the other variables $X_{\neg j}$, and $\widehat{\mathrm{corr}}(X_i, X_j|X_{\neg\{i,j\}})$ is the residual correlation between X_i and X_j after regressing both onto the remaining variables.

We now compare, across three settings, the AUC obtained by either using the absolute value of the regression coefficients $|\widehat{b}_{i\to j}|$ or the absolute value of the corresponding T-statistics $|\widehat{t}_{i\to j}|$ for edge scoring. Results are shown in the left, middle, and right panel of Figure 5.1, respectively.

In the setting with equal error variances $\sigma_i^2 = \sigma_j^2\ \forall i, j$**,** we observe that i) the absolute regression coefficients beat the T-statistics for edge predictions in terms of AUC, and ii) the marginal variances naturally turn out to increase along the causal ordering.

When moving from $|\widehat{b}_{i\to j}|$ to $|\widehat{t}_{i\to j}|$ for scoring edges, we multiply by a term that compares the relative residual variance of X_i and X_j. If X_i is before X_j in the causal ordering it tends to have both smaller marginal and—in our simulation set-up—residual variance than X_j as it becomes increasingly more difficult to predict variables further down the causal ordering. In this case, the fraction of residual variances will tend to be smaller than one and consequently the raw regression coefficients $|\widehat{b}_{i\to j}|$ will be shrunk when moving to $|\widehat{t}_{i\to j}|$. This can explain the worse performance of the T-statistics compared to the raw regression coefficients for edge scoring as scores will tend to be shrunk when in fact $X_i \to X_j$.

Enforcing equal marginal variances by rescaling the rows of B **and the** σ_i^2**'s,** we indeed observe that regression coefficients and T-statistics achieve comparable performance in edge prediction in this somewhat artificial scenario. Here, neither the marginal variances nor the residual variances appear to contain information about the causal ordering any more and the relative ordering between regression coefficients and T-statistics is preserved when multiplying by the factor highlighted in Equation 5.1.

Enforcing decreasing marginal variances by rescaling the rows of B **and the** σ_i^2**'s,** we can, in line with our above reasoning, indeed obtain an artificial scenario in which the T-statistics will outperform the regression coefficients in edge prediction, as now, the factors we multiply by will work in favour of the T-statistics.

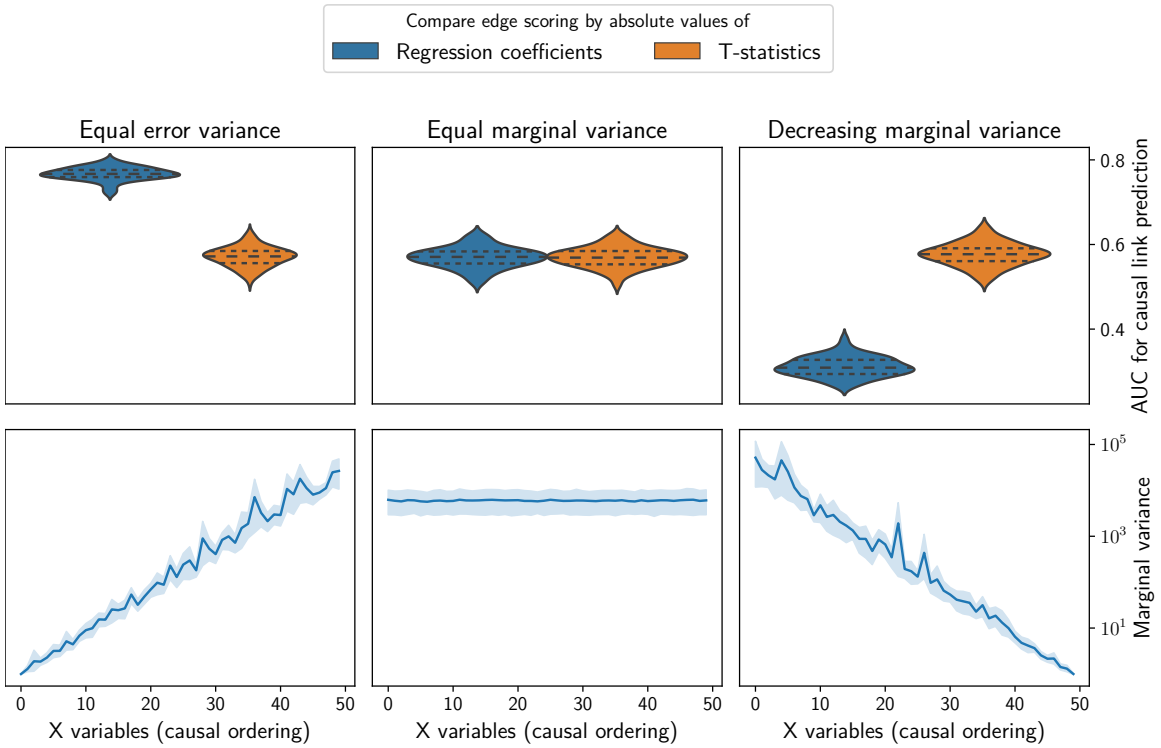


Figure 5.1: Results of the simulation experiment described in Section 5.5.1. Data is generated from an acyclic linear Gaussian model, in turn each variable is regressed onto all remaining variables and either the raw regression coefficient $|\widehat{b}_{i\to j}|$ or the corresponding T-statistics $|\widehat{t}_{i\to j}|$ is used to score the existence of an edge $i \to j$. The top row shows the obtained AUC for causal link prediction and the bottom row the marginal variance of the variables along the causal ordering. The left panel shows naturally increasing marginal variance for equal error variances, for the middle and right panel the model parameters and error variances are rescaled to enforce equal and decreasing marginal variance, respectively.

5.6. Conclusion and Future Work

We believe that competitions like the C4C competition (Runge et al., 2020) and causal discovery benchmark platforms like CauseMe.net (Runge et al., 2019b) are important for bundling and informing the community's joint research efforts into methodology that is readily applicable to tackle real-world data. In practice, there are fundamental limitations to causal structure learning that ultimately require us to employ untestable causal assumptions to proceed towards applications at all. Yet, both these limitations and assumptions are increasingly well understood and characterised by methodological research and time and again need to be challenged and examined through the application to real-world data.

Beyond the algorithms presented here and proposed for baseline benchmarks, different methodology as well as different benchmarks may be of interest. For

example, our methods detect causal links and are viable benchmarks for the structure learning task but they do not per se enable predictions about the interventional distributions.

Acknowledgments

The authors thank Niels Richard Hansen, Steffen Lauritzen, and Jonas Peters for insightful discussions. Thanks to the organisers for a challenging and insightful Causality 4 Climate NeurIPS competition. NT was supported by a research grant (18968) from VILLUM FONDEN. LP and GV were supported by a research grant (13358) from VILLUM FONDEN. MEJ and SW were supported by the Carlsberg Foundation.

Appendices

Appendix A

Distributional Robustness of K-class Estimators and the PULSE

A.1. Structural Equation Models and Interventions

Structural equation models and simultaneous equation models are causal models. That is, they contain more information than the description of an observational distribution. We first introduce the notion of structural equation models (also called structural causal models) and use an example to show how they can be written as in the form of simultaneous equation models (SIM) commonly used in econometrics, see Section A.1.2.

A.1.1. Structural equaton models and interventions

A structural equation model (SEM) (e.g. Bollen, 1989, and Pearl, 2009) over variables $X_1, \dots, X_p$ consists of p assignments of the form

$$X_j := f_j(X_{\mathrm{PA}(j)}, \varepsilon_j), \qquad j = 1, \dots, p,$$

where $\mathrm{PA}(j) \subseteq \{1, \ldots, p\}$ are called the parents of j, together with a distribution over the noise variables $(\varepsilon_1, \ldots, \varepsilon_p)$, which is assumed to have jointly independent marginals. The corresponding graph over $X_1, \ldots, X_p$ is obtained by drawing directed edges from the variables on the right-hand side to the variables on the left-hand side. If the corresponding graph is acyclic, the SEM induces a unique distribution over $(X_1, \ldots, X_p)$, which is often called the observational distribution. Section A.1.2 below discusses an example of linear assignments, which also allows for a cyclic graph structure. The framework of SEMs also models the effect of interventions: An intervention on variable j corresponds to replacing the jth assignment. For example, replacing it by $X_j = 4$, called a hard intervention, or, more generally, by $X_j = g(X_{\widetilde{\mathrm{PA}(j)}}, \tilde{\varepsilon}_j)$ induces yet another distribution over X that is called an interventional distribution and that we denote by $P^{\mathrm{do}(X_j=4)}$ or $P^{\mathrm{do}(X_j=g(X_{\widetilde{\mathrm{PA}(j)}}, \tilde{\varepsilon}_j))}$, respectively. A formal introduction to SEMs, in the general case of cyclic assignments is provided by Bongers et al. (2021), for example. In an SEM, we call all X variables endogenous and, in addition, all variables X_j, for which we have $\mathrm{PA}(j) = \emptyset$, will be called exogenous. A subset of variables is called exogenous relative to another subset if it does not contain a variable that has a parent belonging to the other set.

In the paper, we are mostly interested in one of these equations and we denote the corresponding target variable as Y. Furthermore, some of the other X variables may be unobserved, which we indicate by using the notation H (denoting a vector of variables). In linear models, hidden variables can equivalently be represented as correlation in the noise variables; see e.g. Bongers et al. (2021), and Hyttinen et al. (2012). Finally, we let A denote a collection of variables that are known to enter the system as exogenous variables, relative to (Y, X, H).

A.1.2. Example of a Linear Structural Equation Model

Let the distribution of (Y, X, H, A) be generated according to the possibly cyclic SEM,

$$\begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} := \begin{bmatrix} Y & X^\intercal & H^\intercal \end{bmatrix} B + A^\intercal M + \varepsilon^\intercal. \tag{A.1}$$

Here, B is a square matrix with eigenvalues whose absolute value is strictly smaller than one. This implies that $I - B$ is invertible ensuring that the distribution of (Y, X, H) is well-defined since (Y, X, H) can be expressed in terms of B, M, A and ε as $(I - B^\intercal)^{-1}(M^\intercal A + \varepsilon)$. We denote the random vectors $Y \in \mathbb{R}, X \in \mathbb{R}^d, A \in \mathbb{R}^q, H \in \mathbb{R}^r$ and $\varepsilon \in \mathbb{R}^{d+1+r}$ by target, endogenous regressor, anchor, hidden and noise variables, respectively. We assume that $\varepsilon \perp\!\!\!\perp A$ rendering the so-called anchors as exogenous variables but the coordinate components of A may be dependent on each other. As above, we assume joint independence of the noises $\varepsilon_1, \ldots, \varepsilon_{1+d+r}$. Let $\mathbf{Y} \in \mathbb{R}^{n\times 1}, \mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{A} \in \mathbb{R}^{n\times q}, \mathbf{H} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times(1+d+r)}$ be data-matrices with $n \in \mathbb{N}$ row-wise i.i.d. copies of the variables solving the system in Equation (2.1). Transposing the structural equations and stacking them

vertically by row-wise observations, we can represent all structural equations by

$$[\mathbf{Y}\ \mathbf{X}\ \mathbf{H}] := [\mathbf{Y}\ \mathbf{X}\ \mathbf{H}]B + \mathbf{A}M + \boldsymbol{\varepsilon}.$$

We can solve the structural equations for the endogenous variables and get the so-called structural and reduced form equations, commonly seen in econometrics,

$$\begin{bmatrix}\mathbf{Y} & \mathbf{X} & \mathbf{H}\end{bmatrix}\Gamma = \mathbf{A}M + \boldsymbol{\varepsilon} \quad \text{and} \quad \begin{bmatrix}\mathbf{Y} & \mathbf{X} & \mathbf{H}\end{bmatrix} = \mathbf{A}\Pi + \boldsymbol{\varepsilon}\Gamma^{-1}, \tag{A.2}$$

respectively, where $\Gamma := I - B$ and $\Pi := M\Gamma^{-1}$. Note that the equations in Equation (A.2) differ from the standard representations of simultaneous equation models as we have unobserved endogenous variables $\mathbf{H}$ in the system. In this setup, identifiability of the full system parameters Γ and M in general breaks down due to the dependencies generated by the unobserved endogenous variables. We now assume without loss of generality that Γ has a unity diagonal, such that the target equation of interest, corresponding to the first column of Equation (A.2), is given by

$$\mathbf{Y} = \mathbf{X}\gamma_0 + \mathbf{A}\beta_0 + \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y = \mathbf{Z}\alpha_0 + \tilde{\mathbf{U}}_Y, \tag{A.3}$$

where $(1, -\gamma_0, -\eta_0) \in \mathbb{R}^{(1+d+r)}$, $\beta_0 \in \mathbb{R}^q$ and $\boldsymbol{\varepsilon}_Y$ are the first columns of Γ, M and $\boldsymbol{\varepsilon}$ respectively, $\mathbf{Z} := [\mathbf{X}\ \mathbf{A}]$, $\alpha_0 = (\gamma_0, \beta_0) \in \mathbb{R}^{d+q}$ and $\tilde{\mathbf{U}}_Y := \mathbf{H}\eta_0 + \boldsymbol{\varepsilon}_Y$.

The parameter of interest, α_0, can be derived directly from the corresponding entries in the matrices B and M. It carries causal information in that, for example, after intervening on all variables except for Y, that is, considering an intervention $Z := z$, and $H := h$, Y has the mean $z\alpha_0 + h\eta_0 + E\varepsilon_1$, see Equation (2.1).

In Equation (A.3) we have represented the target variable in terms of a linear combination of the observable variables $Z = (X^\intercal, A^\intercal)^\intercal$ and some unobservable noise term $\tilde{U}_Y$. In contrast to Equation (2.1), Equation (A.3), which is more commonly used in the econometrics literature, models the influence of the latent variables using a dependence between endogenous variables and the noise term $\tilde{U}_Y$; this equivalence is well-known and described by Bongers et al. (2021) and Hyttinen et al. (2012), for example. The construction in Equation (2.1) can be seen as a manifestation of Reichenbach's common cause principle (Reichenbach, 1956). This principle stipulates that if two random variables are dependent then either one causally influences the other or there exists a third variable which causally influences both.

A.2. Algorithms

In this section we present two algorithms. Algorithm 1 details a binary search procedure for the dual PULSE parameter $\lambda_n^\star(p_{\min})$ and Algorithm 2 details the algorithmic construction and output messages of the PULSE estimator.

Algorithm A.1 Binary.Search with precision $1/N$.

input $p_{\min}, N$
if $T_n(\hat{\alpha}^n_{\text{TSLS}}) \geq Q_{\chi^2_q}(1-p_{\min})$ **then** terminate procedure **end if**
$\ell_{\min} \leftarrow 0$; $\ell_{\max} \leftarrow 2$
while $T_n(\hat{\alpha}^n_{\text{K}}(\ell_{\max})) > Q_{\chi^2_q}(1-p_{\min})$ **do**
    $\ell_{\min} \leftarrow \ell_{\max}$; $\ell_{\max} \leftarrow \ell^2_{\max}$
end while
$\Delta \leftarrow \ell_{\max} - \ell_{\min}$
while $\Delta > 1/N$ **do**
    $\ell \leftarrow (\ell_{\min} + \ell_{\max})/2$
    if $T_n(\hat{\alpha}^n_{\text{K}}(\ell)) > Q_{\chi^2_q}(1-p_{\min})$ **then** $\ell_{\min} \leftarrow \ell$ **else** $\ell_{\max} \leftarrow \ell$ **end if**
    $\Delta \leftarrow \ell_{\max} - \ell_{\min}$
end while
return$(\ell_{\max})$

Algorithm A.2 PULSE+

input $p_{\min}$, precision $1/N$, $\hat{\alpha}^n_{\text{ALT}}$
if $T_n(\hat{\alpha}^n_{\text{TSLS}}) \geq Q_{\chi^2_q}(1-p_{\min})$ **then**
    Warning: TSLS outside interior of acceptance region.
    $\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) \leftarrow \hat{\alpha}^n_{\text{ALT}}$
else
    if $T_n(\hat{\alpha}^n_{\text{OLS}}) \leq Q_{\chi^2_q}(1-p_{\min})$ **then**
        Warning: The OLS is accepted.
        $\lambda^\star_n(p_{\min}) \leftarrow 0$
    else
        $\lambda^\star_n(p_{\min}) \leftarrow$ Binary.Search$(N, p_{\min})$
    end if
    $\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) \leftarrow (\mathbf{Z}^\intercal(\mathbf{I} + \lambda^\star_n(p_{\min})P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{I} + \lambda^\star_n(p_{\min})P_{\mathbf{A}})\mathbf{Y}$
end if
return$(\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}))$

A.3. Proofs of Results in Section 2.2

Proof of Proposition 2.1: The minimizations of Equation (2.10) and Equation (2.11) are unconstrained optimization problems. We know that there exists a unique solution if the problems are strictly convex. Thus, it suffices to verify the second order condition for strict convexity of the objective functions, i.e., $D^2 l^n_{\text{K}}(\alpha;\kappa) \succ 0$. To this end, note that $Dl^n_{\text{OLS}}(\alpha;\mathbf{Z}_*,\mathbf{X}) = 2(\alpha^\intercal\mathbf{Z}_*^\intercal\mathbf{Z}_* - \mathbf{Y}^\intercal\mathbf{Z}_*)/n$ and $Dl^n_{\text{IV}}(\alpha;\mathbf{Y},\mathbf{Z}_*,\mathbf{A}) = 2(\alpha^\intercal\mathbf{Z}_*^\intercal P_{\mathbf{A}}\mathbf{Z}_* - \mathbf{Y}^\intercal P_{\mathbf{A}}\mathbf{Z}_*)/n$. Thus, the first order derivative of the K-class regression loss function is given by the κ-weighted affine combination

of these two, that is,

$$\begin{aligned}
&Dl_{\mathrm{K}}^n(\alpha;\kappa,\mathbf{Y},\mathbf{Z}_*,\mathbf{A})\\
&= 2n^{-1}\left((1-\kappa)\left(\alpha^\intercal\mathbf{Z}_*^\intercal\mathbf{Z}_* - \mathbf{Y}^\intercal\mathbf{Z}_*\right) + \kappa\left(\alpha^\intercal\mathbf{Z}_*^\intercal P_{\mathbf{A}}\mathbf{Z}_* - \mathbf{Y}^\intercal P_{\mathbf{A}}\mathbf{Z}_*\right)\right)\\
&= 2n^{-1}\left(\alpha^\intercal\left(\mathbf{Z}_*^\intercal\left((1-\kappa)\mathbf{I} + \kappa P_{\mathbf{A}}\right)\mathbf{Z}_*\right) - \left(\mathbf{Y}^\intercal\left((1-\kappa)\mathbf{I} + \kappa P_{\mathbf{A}}\right)\mathbf{Z}_*\right)\right)\\
&= 2n^{-1}\left(\alpha^\intercal\left(\mathbf{Z}_*^\intercal\left(\mathbf{I} - \kappa\left(\mathbf{I} - P_{\mathbf{A}}\right)\right)\mathbf{Z}_*\right) - \left(\mathbf{Y}^\intercal\left(\mathbf{I} - \kappa\left(\mathbf{I} - P_{\mathbf{A}}\right)\right)\mathbf{Z}_*\right)\right)\\
&= 2n^{-1}\left(\alpha^\intercal\left(\mathbf{Z}_*^\intercal\left(\mathbf{I} - \kappa P_{\mathbf{A}}^\perp\right)\mathbf{Z}_*\right) - \left(\mathbf{Y}^\intercal\left(\mathbf{I} - \kappa P_{\mathbf{A}}^\perp\right)\mathbf{Z}_*\right)\right),
\end{aligned}$$

where $P_{\mathbf{A}}^\perp = \mathbf{I} - P_{\mathbf{A}}$. The second order derivative is given by

$$D^2 l_{\mathrm{K}}^n(\alpha;\kappa,\mathbf{Y},\mathbf{Z}_*,\mathbf{A}) = 2n^{-1}\mathbf{Z}_*^\intercal(I - \kappa P_{\mathbf{A}}^\perp)\mathbf{Z}_*,$$

The second derivative is and is proportional to the matrix we need to invert in order to solve the normal equation that yields the K-class estimator. As a consequence, we have that the K-class estimator is guaranteed to exist and be unique if the second derivative is strictly positive definite, i.e., invertible.

Let us first consider $\kappa < 1$. To see that $D^2 l_{\mathrm{K}}^n(\alpha;\kappa,\mathbf{Y},\mathbf{Z}_*,\mathbf{A}) \succ 0$, take any $y \in \mathbb{R}^{d_1+q_1} \setminus \{0\}$ and assume that Assumption 2.2.(a) holds. That is, we assume that $\operatorname{rank}(\mathbf{Z}_*^\intercal\mathbf{Z}_*) = \operatorname{rank}(\mathbf{Z}_*) = d_1 + q_1$ almost surely such that $z = \mathbf{Z}_* y \in \mathbb{R}^n \setminus \{0\}$ almost surely. Without Assumption 2.2.(a), choosing $y \in \ker(\mathbf{Z}_*) \setminus \{0\}$ yields a zero in the following quadratic form with positive probability. However, with this assumption (disregarding $2n^{-1}$) we get that

$$\begin{aligned}
y^\intercal D^2 l_{\mathrm{K}}^n(\alpha;\kappa)y \propto (1-\kappa)y^\intercal\mathbf{Z}_*^\intercal\mathbf{Z}_* y + \kappa y^\intercal\mathbf{Z}_*^\intercal P_{\mathbf{A}}\mathbf{Z}_* y &= (1-\kappa)\|z\|_2^2 + \kappa\|P_{\mathbf{A}}z\|_2^2\\
&\geq \begin{cases} (1-\kappa)\|z\|_2^2 + \kappa\|z\|_2^2 = \|z\|_2^2, & \text{if } \kappa \in (-\infty, 0),\\ (1-\kappa)\|z\|_2^2, & \text{if } \kappa \in [0,1), \end{cases} > 0.
\end{aligned}$$

Here, we used that $P_{\mathbf{A}} = P_{\mathbf{A}}^\intercal = \mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal$ is an orthogonal projection matrix, hence $P_{\mathbf{A}} = P_{\mathbf{A}}^\intercal P_{\mathbf{A}}$ and $0 \leq \|P_{\mathbf{A}}w\|_2^2 \leq \|w\|_2^2$ for any $w \in \mathbb{R}^q$.

Let us now consider the case $\kappa = 1$. The quadratic form is now given by

$$y^\intercal D^2 l_{\mathrm{K}}^n(\alpha;\kappa)y = \|P_{\mathbf{A}}z\|_2^2 = y^\intercal\mathbf{Z}_*^\intercal\mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{Z}_* y.$$

If $\operatorname{rank}(\mathbf{A}^\intercal\mathbf{Z}_*) < d_1 + q_1$ with positive probability, then any $y \in \ker(\mathbf{A}^\intercal\mathbf{Z}_*)\setminus\{0\} \neq \emptyset$ yields a zero quadratic value, showing that $l_{\mathrm{K}}^n(\alpha;\kappa)$ is not strictly convex with positive probability. However, if Assumption 2.2.(b) holds, i.e., that $\mathbf{A}^\intercal\mathbf{Z}_* \in \mathbb{R}^{q\times(d_1+q_1)}$ satisfies $\operatorname{rank}(\mathbf{A}^\intercal\mathbf{Z}_*) = d_1 + q_1$ almost surely, then $D^2 l_{\mathrm{K}}^n(\alpha;\kappa)$ is also guaranteed to be positive definite almost surely.

Thus, we have shown sufficient conditions for $D^2 l_{\mathrm{K}}^n(\alpha;\kappa)$ to be almost surely positive definite, ensuring strict convexity of the $l_{\mathrm{K}}^n(\alpha;\kappa)$, hence almost sure uniqueness of a global minimum. The unique global minimum is then found as a solution to the normal equation $Dl_{\mathrm{K}}^n(\alpha;\kappa) = 0$ which is given by $\hat{\alpha}_{\mathrm{K}}^n(\kappa) = (\mathbf{Z}_*^\intercal(\mathbf{I} - \kappa P_{\mathbf{A}}^\perp)\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\intercal(\mathbf{I} - \kappa P_{\mathbf{A}}^\perp)\mathbf{Y}$. We conclude that under the above conditions the K-class estimator $\hat{\alpha}_{\mathrm{K}}^n(\kappa)$ solves the unconstrained minimization problem $\arg\min_{\alpha\in\mathbb{R}^{d_1+q_1}} l_{\mathrm{K}}^n(\alpha;\kappa)$ almost surely. □

Proof of Proposition 2.2: We first prove that the population estimand that minimizes the population loss function is well-defined. It suffices to show strict convexity of the population loss function. Let Assumption 2.3.(a) hold, i.e., that $\mathrm{Var}(Z_*)$ is positive definite, and consider $\kappa \in [0, 1)$. For any $y \in \mathbb{R}^{d_1+q_1} \setminus \{0\}$ we see that

$$\begin{aligned} y^\intercal D^2 l_{\mathrm{K}}(\alpha; \kappa) y &= (1-\kappa) y^\intercal E(Z_* Z_*^\intercal) y + \kappa y^\intercal E(Z_* A^\intercal) E(AA^\intercal)^{-1} E(A Z_*^\intercal) y \\ &\geq (1-\kappa) y^\intercal E(Z_* Z_*^\intercal) y \\ &= (1-\kappa) \left(y^\intercal \mathrm{Var}(Z_*) y + y^\intercal E(Z_*) E(Z_*)^\intercal y \right) \\ &\geq (1-\kappa) y^\intercal \mathrm{Var}(Z_*) y > 0, \end{aligned} \tag{A.4}$$

proving strict convexity of the K-class penalized loss function. Now let $\kappa = 1$ and let Assumption 2.1.(h) and Assumption 2.3.(b) hold, i.e., $\mathrm{Var}(A)$ is positive definite and $E(AZ_*^\intercal)$ is of full column rank (which implicitly assumes we are in the just- or over-identified case). First note that by the above considerations this implies that $E(AA^\intercal)$ and its inverse $E(AA^\intercal)^{-1}$ are positive definite. For any $y \in \mathbb{R}^{d_1+q_1} \setminus \{0\}$ we note that $z := E(AZ_*^\intercal) y \neq 0$ by injectivity of $E(AZ_*^\intercal)$, and hence

$$y^\intercal D^2 l_{\mathrm{K}}(\alpha; \kappa) y = z^\intercal E(AA^\intercal)^{-1} z > 0,$$

by the positive definiteness of $E(AA^\intercal)^{-1}$. Proving strict convexity.

In both setups the minimization estimator of the population loss function solves the normal equation $0 = D l_{\mathrm{K}}(\alpha; \kappa) = (1-\kappa) D l_{\mathrm{OLS}}(\alpha) + \kappa D l_{\mathrm{IV}}(\alpha)$ which by rearranging the terms yields that

$$\begin{aligned} \alpha_{\mathrm{K}}(\kappa) = &\Big((1-\kappa) E(Z_* Z_*^\intercal) + \kappa E(Z_* A^\intercal) E(AA^\intercal)^{-1} E(A Z_*^\intercal) \Big)^{-1} \\ &\cdot \Big((1-\kappa) E(Z_* Y) + \kappa E(Z_* A^\intercal) E(AA^\intercal)^{-1} E(AY) \Big). \end{aligned}$$

We now prove that the estimators are asymptotically well-defined if the population conditions of Assumption 2.3.(a) and Assumption 2.3.(b) hold. For $\kappa \in [0, 1)$, we know from Proposition 2.1 that

$$\begin{aligned} &P\left[\operatorname*{arg\,min}_{\alpha \in \mathbb{R}^{d_1+q_1}} l_{\mathrm{K}}^n(\alpha; \kappa, \mathbf{Y}, \mathbf{Z}_*, \mathbf{A}) \text{ is well-defined} \right] \\ &\geq P\left[\mathbf{Z}_*^\intercal \mathbf{Z}_* \text{ is positive definite} \right], \end{aligned}$$

So it suffices to show that the lower converges to one in probability. By the weak law of large numbers we have, for any $\varepsilon > 0$ that $P(\|\mathbf{Z}_*^\intercal \mathbf{Z}_* - E(Z_* Z_*^\intercal)\| < \varepsilon) \to 1$. Note that by Assumption 2.3.(a), i.e., that $\mathrm{Var}(Z_*)$ is positive definite, we also have that $E(Z_* Z_*^\intercal)$ is positive definite; see Equation (A.4) above. Note that the set of positive definite matrices S_+ is an open set in the space of symmetric matrices S of the same dimensions. Hence, there must exist an open ball $B(E(Z_* Z_*^\intercal), c) \subseteq S_+$ with center $E(Z_* Z_*^\intercal)$ and radius $c > 0$, fully contained in the set of positive definite matrices. By virtue of the above convergence in probability, we have that

$$\begin{aligned} P\left[\mathbf{Z}_*^\intercal \mathbf{Z}_* \text{ is positive definite} \right] &\geq P\left(\mathbf{Z}_*^\intercal \mathbf{Z}_* \in B(E(Z_* Z_*^\intercal), c) \right) \\ &\geq P(\|\mathbf{Z}_*^\intercal \mathbf{Z}_* - E(Z_* Z_*^\intercal)\| < c) \to 1, \end{aligned}$$

proving that the estimator minimizing the K-class penalized regression function is asymptotically well-defined. In the case of $\kappa = 1$ the argument for asymptotic well-definedness follows by almost the same arguments. Arguing that $\mathbf{A}^\intercal\mathbf{A}$ is positive definite with probability converging to one since $\mathrm{Var}(A)$ is assumed positive definite follows from the same arguments as above. To see that $\mathbf{A}^\intercal\mathbf{Z}_*$ is of full column rank with probability converging to one, we use that $E(AZ_*^\intercal)$ is assumed full column rank. If $q = d_1 + q_2$, then follows from the above arguments. Otherwise, if $q > d_1 + q_1$, then we modify the above arguments using that the set of injective linear maps from $\mathbb{R}^{d_1+q_1}$ to $\mathbb{R}^q$ is an open set of all linear maps from $\mathbb{R}^{d_1+q_1}$ to $\mathbb{R}^q$.

Finally, by the law of large numbers, Slutsky's theorem and the continuous mapping theorem, one can easily realize that $\hat{\alpha}_{\mathrm{K}}^n(\kappa) \xrightarrow{P} \alpha_{\mathrm{K}}(\kappa)$. □

Proof of Theorem 2.1: Let (Y, X, H, A) be generated by the SEM given by

$$[Y\ X^\intercal\ H^\intercal]^\intercal := B[Y\ X^\intercal\ H^\intercal]^\intercal + MA + \varepsilon, \tag{A.5}$$

where ε satisfies $\varepsilon \perp\!\!\!\perp A$ and has jointly independent marginals $\varepsilon_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp \varepsilon_{d+1+r}$ with finite second moment $E\|\varepsilon\|_2^2 < \infty$ and mean zero $E(\varepsilon) = 0$. The distribution of A is determined independently of Equation (A.5) and with the only requirement that $E\|A\|_2^2 < \infty$. Note that we have transposed B and M for ease of notation. This implies that (Y, X, H) satisfies the reduced form equations given by $[Y\ X^\intercal\ H^\intercal]^\intercal = \Pi A + \Gamma^{-1}\varepsilon$, where $\Gamma = I - B$ and $\Pi = \Gamma^{-1}M$.

Now let $X_* \subseteq X$ and $A_* \subseteq A$ be our candidate predictors of Y, regardless of which variables directly affect Y and let $Z_* = [X_*^\intercal\ A_*^\intercal]^\intercal$. By the reduced form structural equations we derive the marginal reduced forms as

$$Y = \Pi_Y A + \Gamma_Y^{-1}\varepsilon \quad \text{and} \quad X_* = \Pi_{X_*} A + \Gamma_{X_*}^{-1}\varepsilon, \tag{A.6}$$

where $\Pi_Y, \Pi_{X_*}, \Gamma_Y^{-1}, \Gamma_{X_*}^{-1}$ are the relevant sub-matrices of rows from Π and Γ^{-1}. Furthermore, let (Y^v, X^v, H^v) be generated as a solution to the SEM of Equation (A.5) under the intervention $\mathrm{do}(A := v)$, where $v \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ is any fixed stochastic element uncorrelated with ε. Under the intervention and by similar manipulations as above, we arrive at the following marginal reduced forms $Y^v = \Pi_Y v + \Gamma_Y^{-1}\varepsilon$ and $X_*^v = \Pi_{X_*} v + \Gamma_{X_*}^{-1}\varepsilon$. For a fixed γ and β, with A_{-*} being $A \setminus A_*$, we have that

$$\begin{aligned} Y - \gamma^\intercal X_* - \beta^\intercal A_* &= (\Pi_Y - \gamma^\intercal \Pi_{X_*})A + (\Gamma_Y^{-1} - \Gamma_{X_*}^{-1})\varepsilon - \beta^\intercal A_* \\ &= (\delta_1^\intercal - \beta^\intercal)A_* + \delta_2^\intercal A_{-*} + w^\intercal\varepsilon = \xi^\intercal A + w^\intercal\varepsilon, \end{aligned}$$

where δ_1, δ_2 are such that $(\Pi_Y - \gamma^\intercal\Pi_{X_*})A = \delta_1^\intercal A_* + \delta_2^\intercal A_{-*}$, ξ is such that $\xi^\intercal A = (\delta_1^\intercal - \beta^\intercal)A_* + \delta_2^\intercal A_{-*}$ and $w^\intercal := (\Gamma_Y^{-1} - \Gamma_{X_*}^{-1})$. Similar manipulations yield that the regression residuals under the intervention are given by $Y^v - \gamma^\intercal X_*^v - \beta^\intercal v_* = \xi^\intercal v + w^\intercal\varepsilon$. Since $A \perp\!\!\!\perp \varepsilon$ and ε has mean zero, we have that

$$E(Y - \gamma^\intercal X_* - \beta^\intercal A_*|A) = \xi^\intercal A + w^\intercal E(\varepsilon) = \xi^\intercal A, \tag{A.7}$$

$$Y - \gamma^\intercal X_* - \beta^\intercal A_* - E(Y - \gamma^\intercal X_* - \beta^\intercal A_*|A) = w^\intercal\varepsilon. \tag{A.8}$$

By construction $E(v\varepsilon^\intercal) = 0$, so

$$\begin{aligned} E^{\mathrm{do}(A:=v)}\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)^2\right] &= E\left[(\xi^\intercal v + w^\intercal \varepsilon)^2\right] \\ &= E\left[(\xi^\intercal v)^2\right] + E\left[(w^\intercal \varepsilon)^2\right] + \xi^\intercal E(v\varepsilon^\intercal) w \\ &= E\left[(\xi^\intercal v)^2\right] + E\left[(w^\intercal \varepsilon)^2\right]. \end{aligned} \tag{A.9}$$

We investigate the terms of Equation (A.9) and note by Equation (A.8) that

$$\begin{aligned} E\left[(w^\intercal \varepsilon)^2\right] &= E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_* - E(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A))^2\right] \\ &= E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)^2\right] + E\left[E\left(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A\right)^2\right] \\ &\quad - 2E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)\, E(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A)\right]. \end{aligned} \tag{A.10}$$

In Equation (A.7) we established that $E(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A)$ is a linear function of A, so it must hold that

$$\begin{aligned} E(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A) &= \operatorname*{arg\,min}_{Z \in \sigma(A)} \|Y - \gamma^\intercal X_* - \beta^\intercal A_* - Z\|^2_{L^2(P)} \\ &= A^\intercal \operatorname*{arg\,min}_{c \in \mathbb{R}^q} \|Y - \gamma^\intercal X_* - \beta^\intercal A_* - A^\intercal c\|^2_{L^2(P)} \\ &= A^\intercal E(AA^\intercal)^{-1} E\left[A\left(Y - \gamma^\intercal X_* - \beta^\intercal A_*\right)\right], \end{aligned}$$

almost surely. In the first equality we used that the conditional expectation is the best predictor under the $L^2(P)$-norm and in the third equality we used that the minimizer is given by the population ordinary least square estimate. An immediate consequence of this is that the second term of Equation (A.10) equals

$$\begin{aligned} E\left[E\left(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A\right)^2\right] = E[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)A^\intercal] E(AA^\intercal)^{-1} \\ \cdot E[A(Y - \gamma^\intercal X_* - \beta^\intercal A_*)], \end{aligned}$$

which is seen to be of the same form of the third term in Equation (A.10),

$$\begin{aligned} &E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)\, E(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A)\right] \\ =&E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)\, A^\intercal\right] E(AA^\intercal)^{-1} E\left[A(Y - \gamma^\intercal X_* - \beta^\intercal A_*)\right]. \end{aligned}$$

Thus, we conclude that the second term of Equation (A.9) is given by

$$\begin{aligned} E\left[(w^\intercal \varepsilon)^2\right] &= E\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)^2\right] - E\left[E\left(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A\right)^2\right] \\ &= l_{\mathrm{OLS}}(\alpha; Y, Z_*) - l_{\mathrm{IV}}(\alpha; Y, Z_*, A). \end{aligned}$$

Taking the supremum over all $v \in C(\kappa)$ of the first term of Equation (A.9) we

obtain

$$\begin{aligned}\sup_{v \in C(\kappa)} E\left[(\xi^\intercal v)^2\right] &= \sup_{v \in C(\kappa)} \xi^\intercal E\left[vv^\intercal\right]\xi \\ &= \frac{1}{1-\kappa}\xi^\intercal E\left[AA^\intercal\right]\xi \\ &= \frac{1}{1-\kappa}E\left[(\xi^\intercal A)^2\right] \\ &= \frac{1}{1-\kappa}E\left[E\left(Y - \gamma^\intercal X_* - \beta^\intercal A_* | A\right)^2\right] \\ &= \frac{1}{1-\kappa} l_{\mathrm{IV}}(\alpha; Y, Z_*, A),\end{aligned}$$

where the second last equation follows from Equation (A.7) and the second equation follows from the following argument. For any $v \in C(\kappa)$ we have that $E(vv^\intercal) \preceq \frac{1}{1-\kappa}E(AA^\intercal)$, that is, for all $x \in \mathbb{R}^q$ it holds that $\frac{1}{1-\kappa}x^\intercal E(AA^\intercal)x \geq x^\intercal E(vv^\intercal)x$, which implies that the upper bound is attained for any v such that $E(vv^\intercal) = \frac{1}{1-\kappa}E(AA^\intercal)$. Thus, we have that

$$\begin{aligned}&\sup_{v \in C(\kappa)} E^{\mathrm{do}(A:=v)}\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)^2\right] \\ &= \sup_{v \in C(\kappa)} E\left[(\xi^\intercal v)^2\right] + E\left[(w^\intercal \varepsilon)^2\right] \\ &= l_{\mathrm{OLS}}(\alpha; Y, Z_*) + \frac{\kappa}{1-\kappa} l_{\mathrm{IV}}(\alpha; Y, Z_*, A).\end{aligned}$$

By the representation in Equation (2.13) it therefore follows that the population K-class estimate with parameter $\kappa \neq 1$ is given as the estimate that minimizes the worst case mean squared prediction error over all interventions contained in $C(\kappa)$, that is,

$$\alpha_{\mathrm{K}}(\kappa; Z_*, A) = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^d, \beta \in \mathbb{R}^{q_1}} \sup_{v \in C(\kappa)} E^{\mathrm{do}(A:=v)}\left[(Y - \gamma^\intercal X_* - \beta^\intercal A_*)^2\right].$$

□

A.4. Proofs of Selected Results in Section 2.3

Corollary A.1 (K-class estimators differ)**.** *Let Assumptions 2.6 and 2.9 hold. If* $\lambda_1, \lambda_2 \geq 0$ *with* $\lambda_1 \neq \lambda_2$*, then* $\hat{\alpha}^n_{\mathrm{K}}(\lambda_1) \neq \hat{\alpha}^n_{\mathrm{K}}(\lambda_2)$*.*

Proof of Corollary A.1: Let Assumptions 2.6 and 2.9 hold. $\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is well-defined for all $\lambda \geq 0$ by Proposition 2.1. Let $\lambda_1, \lambda_2 \geq 0$ with $\lambda_1 \neq \lambda_2$ and note that the orthogonality condition derived in the proof of Lemma 2.3 also applies here. That

is, $\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_i), (\mathbf{I} + \lambda_i P_{\mathbf{A}})z\rangle = 0$, for all $z \in \mathcal{R}(\mathbf{Z})$ and $i = 1, 2$. Assume for contradiction that $\hat{\alpha}_{\mathrm{K}}^n(\lambda_1) = \hat{\alpha}_{\mathrm{K}}^n(\lambda_2)$. This implies that

$$\begin{aligned} 0 &= \langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1), (\mathbf{I} + \lambda_1 P_{\mathbf{A}})z - (\mathbf{I} + \lambda_2 P_{\mathbf{A}})z\rangle \\ &= \langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1), (\lambda_1 - \lambda_2)P_{\mathbf{A}}z\rangle = (\lambda_1 - \lambda_2)\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1), P_{\mathbf{A}}z\rangle, \end{aligned}$$

for any $z \in \mathcal{R}(\mathbf{Z})$. Thus, by symmetry and idempotency of $P_{\mathbf{A}}$ we have that for all $z \in \mathcal{R}(\mathbf{Z})$,

$$\langle P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1), P_{\mathbf{A}}z\rangle = \langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1), P_{\mathbf{A}}z\rangle = 0.$$

That is, $P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)$ is the orthogonal projection of $P_{\mathbf{A}}\mathbf{Y}$ onto $\mathcal{R}(P_{\mathbf{A}}\mathbf{Z})$. This is equivalent with saying that $\hat{\alpha}_{\mathrm{K}}^n(\lambda_1) \in \mathcal{M}_{\mathrm{IV}}$ as the space of minimizers of l_{IV}^n are exactly the coefficients in $\mathbb{R}^{d_1+q_1}$ which mapped through $P_{\mathbf{A}}\mathbf{Z}$ yields this orthogonal projection. See the proof of Lemma 2.3 for further elaboration on this equivalence. This is a contradiction to Assumption 2.9, hence $\hat{\alpha}_{\mathrm{K}}^n(\lambda_1) \neq \hat{\alpha}_{\mathrm{K}}^n(\lambda_2)$. □

Lemma A.1 (Monotonicity of the losses and the test statistic)**.**
When Assumption 2.6.(a) holds the maps $[0, \infty) \ni \lambda \mapsto l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *and* $[0, \infty) \ni \lambda \mapsto l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *are monotonically increasing and monotonically decreasing, respectively. Consequently, if Assumption 2.7 holds, we have that the map* $[0, \infty) \ni \lambda \longmapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ *is monotonically decreasing. Furthermore, if Assumption 2.9 also holds, these monotonicity statements can be strengthened to strictly decreasing and strictly increasing.*

Proof of Lemma A.1: Let Assumption 2.6.(a) hold, such that $\hat{\alpha}_{\mathrm{K}}^n(\lambda)$ is well-defined for all $\lambda \geq 0$; see Proposition 2.1. Let $\lambda_2 > \lambda_1 \geq 0$ and note that

$$\begin{aligned} l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) + \lambda_1 l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) &\leq l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)) + \lambda_1 l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)) \\ &= l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)) + \lambda_2 l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)) + (\lambda_1 - \lambda_2) l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)) \\ &\leq l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) + \lambda_2 l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) + (\lambda_1 - \lambda_2) l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)), \end{aligned}$$

where we used that $\hat{\alpha}_{\mathrm{K}}^n(\lambda)$ minimizes the expressions with penalty factor λ. Thus,

$$(\lambda_1 - \lambda_2) l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) \leq (\lambda_1 - \lambda_2) l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)),$$

which is equivalent with

$$l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) \geq l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)),$$

proving that $\lambda \mapsto l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ is monotonically decreasing.

If $\lambda_2 > \lambda_1 = 0$, then we note that

$$l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_1)) = \min_{\alpha}\{l_{\mathrm{OLS}}^n(\alpha)\} \leq l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_2)).$$

For any $\lambda > 0$,

$$\hat{\alpha}_{\mathrm{K}}^n(\lambda) = \arg\min_{\alpha}\{l_{\mathrm{OLS}}^n(\alpha) + \lambda l_{\mathrm{IV}}^n(\alpha)\} = \arg\min_{\alpha}\{\lambda^{-1} l_{\mathrm{OLS}}^n(\alpha) + l_{\mathrm{IV}}^n(\alpha)\}.$$

Thus, if $\lambda_2 > \lambda_1 > 0$, we have that

$$\begin{aligned}
&\lambda_1^{-1} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_1)) + l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_1)) \\
\leq\, &\lambda_1^{-1} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)) + l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)) \\
=\, &\lambda_2^{-1} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)) + l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)) + (\lambda_1^{-1} - \lambda_2^{-1}) l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)) \\
\leq\, &\lambda_2^{-1} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_1)) + l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_1)) + (\lambda_1^{-1} - \lambda_2^{-1}) l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2)),
\end{aligned}$$

hence $l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_1)) \leq l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda_2))$, so $\lambda \mapsto l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ is monotonically increasing.

When Assumption 2.7 holds, the map

$$\lambda \mapsto T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) = n \frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))},$$

is well-defined and monotonically decreasing, as it is given by a positive, monotonically decreasing function over a strictly positive and monotonically increasing function.

Furthermore, when Assumption 2.9 holds, Corollary 2.1 yields that for $\lambda_1, \lambda_2 \geq 0$ with $\lambda_1 \neq \lambda_2$ it holds that $\hat{\alpha}^n_{\mathrm{K}}(\lambda_1) \neq \hat{\alpha}^n_{\mathrm{K}}(\lambda_2)$. As a consequence, the above inequalities become strict, since otherwise (Dual.$\lambda.n$) has two distinct solutions which contradicts Proposition 2.1. Replacing the above inequalities with strict inequalities yields that the functions are strictly increasing and decreasing, respectively. □

Lemma A.2. *Let $p_{\min} \in (0,1)$ and let Assumption 2.6.(a) and Assumption 2.7 hold. If $\lambda^\star_n(p_{\min}) < \infty$, it holds that*

$$T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) \leq Q_{\chi^2_q}(1 - p_{\min}). \tag{A.11}$$

If the ordinary least square estimator satisfies $T_n(\hat{\alpha}^n_{\mathrm{OLS}}) < Q_{\chi^2_q}(1 - p_{\min})$, then Equation (A.11) *holds with strict inequality, otherwise it holds with equality.*

Proof of Lemma A.2: Let $p_{\min} \in (0,1)$ and let let Assumption 2.6.(a) and Assumption 2.7 hold, such that $\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ for all $\lambda \geq 0$ and $T_n(\alpha)$ for all $\alpha \in \mathbb{R}^{d_1+q_1}$ are well-defined, by Proposition 2.1.

Assume that $\lambda^\star_n(p_{\min}) < \infty$, so we know that $T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q_{\chi^2_q}(1-p)$ for all $\lambda > \lambda^\star_n(p_{\min})$ by the monotonicity of Lemma 2.6. Thus, the first statement follows if we can show that $\lambda \mapsto T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ is a continuous function. Since $\alpha \mapsto T_n(\alpha)$ is continuous it suffices to show that $[0,\infty) \ni \lambda \mapsto \hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is continuous. Recall that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = (\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Y}$, for any $\lambda \geq 0$. Note that the functions $\mathrm{Inv} : S^{d_1+q_1}_{++} \to S^{d_1+q_1}_{++}$ given by $\mathbf{M} \overset{\mathrm{Inv}}{\mapsto} \mathbf{M}^{-1}$, $\lambda \mapsto \mathbf{Z}^\intercal(\mathbf{I}+\lambda P_{\mathbf{A}})\mathbf{Z}$, $\lambda \mapsto \mathbf{Z}^\intercal(\mathbf{I}+\lambda P_{\mathbf{A}})\mathbf{Y}$ and $(\mathbf{B},\mathbf{C}) \mapsto \mathbf{BC}$ are all continuous maps, where $S^{d_1+q_1}_{++}$ is the set of all positive definite $(d_1+q_1)\times(d_1+q_1)$ matrices. We have that $\lambda \mapsto \hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is a composition of these continuous maps, hence it itself is continuous. This proves the first statement.

In the case that OLS is strictly feasible in the PULSE problem, $T_n(\hat{\alpha}^n_{\mathrm{OLS}}) < Q_{\chi^2_q}(1-p_{\min})$, we have that

$$\lambda^\star(p_{\min}) = \inf\left\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q_{\chi^2_q}(1-p_{\min})\right\} = 0,$$

since $\hat{\alpha}^n_{\mathrm{K}}(0) = \hat{\alpha}^n_{\mathrm{OLS}}$, hence

$$T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) = T_n(\hat{\alpha}^n_{\mathrm{K}}(0)) = T_n(\hat{\alpha}^n_{\mathrm{OLS}}) < Q_{\chi^2_q}(1-p_{\min}).$$

Similar arguments show that, if the OLS is just-feasible in the PULSE problem, $T_n(\hat{\alpha}^n_{\mathrm{OLS}}) = Q_{\chi^2_q}(1-p_{\min})$, then $T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) = Q_{\chi^2_q}(1-p_{\min})$.

In the case that the OLS estimator is infeasible in the PULSE problem, $Q_{\chi^2_q}(1-p_{\min}) < T_n(\hat{\alpha}^n_{\mathrm{OLS}})$, continuity and monotonicity of $\lambda \mapsto T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ entail it must hold that $T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) = Q_{\chi^2_q}(1-p_{\min})$, as otherwise

$$T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) < Q_{\chi^2_q}(1-p_{\min}) < T_n(\hat{\alpha}^n_{\mathrm{K}}(0)),$$

implying that there exists $\tilde{\lambda} < \lambda^\star_n(p_{\min})$ such that $T_n(\hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda})) \leq Q_{\chi^2_q}(1-p_{\min})$, contradicting $\lambda^\star_n(p_{\min}) = \inf\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q_{\chi^2_q}(1-p_{\min})\}$. □

A.5. Proofs of Remaining Results in Section 2.3

Proof of Lemma 2.1: We want to show an asymptotic guarantee that type I errors (rejecting a true hypothesis) occur with probability p. That is, if $\mathcal{H}_0(\alpha)$ is true, then $P(T^c_n(\alpha) > Q_{\chi^2_q}(1-p)) \overset{n\to\infty}{\longrightarrow} p$. Furthermore, we want to show that for any fixed alternative, the probability of type II errors (failure to reject) converges to zero. That is, if P is such that $\mathcal{H}_0(\alpha)$ is false, then $P(T^c_n(\alpha) \leq Q_{\chi^2_q}(1-p)) \overset{n\to\infty}{\longrightarrow} 0$.

Fix any $\alpha \in \mathbb{R}^{d_1+q_1}$. It suffices to show that under the null-hypothesis $T^c_n(\alpha)$ is asymptotically Chi-squared distributed with q degrees of freedom and that $T^c_n(\alpha)$ tends to infinity under any fixed alternative. Without loss of generality assume that $c(n) = n$ for all $n \in \mathbb{N}$ and recall that

$$T^c_n(n) = T_n(\alpha) = n\frac{l^n_{\mathrm{IV}}(\alpha)}{l^n_{\mathrm{OLS}}(\alpha)} = n\frac{\|P_{\mathbf{A}}(\mathbf{Y}-\mathbf{Z}\alpha)\|_2^2}{\|\mathbf{Y}-\mathbf{Z}\alpha\|_2^2}.$$

By the idempotency of $P_{\mathbf{A}}$ the numerator can be rewritten as

$$\|P_{\mathbf{A}}(\mathbf{Y}-\mathbf{Z}\alpha)\|_2^2 = \|(\mathbf{A}^\intercal\mathbf{A})^{-1/2}\mathbf{A}^\intercal\mathbf{R}(\alpha)\|_2^2,$$

while the denominator takes the form $\|\mathbf{R}(\alpha)\|_2^2$. Here, $\mathbf{R}(\alpha) := \mathbf{Y}-\mathbf{Z}\alpha$ and $R(\alpha) := Y - Z^\intercal\alpha$ denotes the empirical and population regression residuals, respectively. Assumption 2.7 ensures that T_n is well-defined on the entire domain of $\mathbb{R}^{d_1+q_1}$ as the denominator is never zero. Furthermore, note that both $R(\alpha)$ for

any $\alpha \in \mathbb{R}^{d_1+q_1}$ and A_i for any $i = 1, ..., q$ have finite second moments by virtue of Assumption 2.1.(f).

Assume that the null hypothesis of zero correlation between the components of A and the regression residuals $R(\alpha)$ holds. First we show that the null hypothesis, under the stated assumptions, implies independence between the exogenous variables A and the regression residuals $R(\alpha)$. It holds that $E(AR(\alpha)) = E(A)E(R(\alpha)) = 0$ by Assumption 2.4.(b), the mean zero assumption of A. Assumption 2.4.(a), i.e., $A \perp\!\!\!\perp U_Y$, yields that

$$0 = E(AR(\alpha)) = E(AZ^\intercal)(\alpha_0 - \alpha) + E(AU_Y) = E(AZ^\intercal)(\alpha_0 - \alpha), \qquad \text{(A.12)}$$

proving that $\alpha - \alpha_0 = w$ for some $w \in \text{kern}(E(AZ^\intercal))$. Recall that the marginal structural equation of Equation (A.6) states that $X_* = \Pi_{X_*} A + \Gamma_{X_*}^{-1}\varepsilon$. Thus, Z has the following representation

$$\begin{aligned} Z = \begin{bmatrix} X_* \\ A_* \end{bmatrix} &= \begin{bmatrix} \Pi_{X_*} A + \Gamma_{X_*}^{-1}\varepsilon \\ A_* \end{bmatrix} \\ &= \begin{bmatrix} \Pi_{X_*}^{(*)} & \Pi_{X_*}^{(-*)} \\ I & 0 \end{bmatrix} \begin{bmatrix} A_* \\ A_{-*} \end{bmatrix} + \begin{bmatrix} \Gamma_{X_*}^{-1} \\ 0 \end{bmatrix} \varepsilon =: \Lambda A + \Psi\varepsilon, \end{aligned}$$

where $\Pi_{X_*} = [\Pi_{X_*}^{(*)} \; \Pi_{X_*}^{(-*)}] \in \mathbb{R}^{d_1 \times (q_1+q_2)}$ and Λ, Ψ are the conformable block-matrices. Since $A \perp\!\!\!\perp \varepsilon$ by Assumption 2.1.(d) we have that $E(A\varepsilon^\intercal) = 0$, hence

$$0 = E(AZ^\intercal)w = E(AA^\intercal)\Lambda^\intercal w + E(A\varepsilon^\intercal)\Psi^\intercal w = E(AA^\intercal)\Lambda^\intercal w.$$

This proves that $\Lambda^\intercal w = 0$ as $E(AA^\intercal)$ is of full rank by Assumption 2.1.(h). Hence,

$$\begin{aligned} R(\alpha) = Y - Z^\intercal\alpha &= Z^\intercal(\alpha_0 - \alpha) + U_Y = Z^\intercal w + U_Y \\ &= A^\intercal\Lambda^\intercal w + \varepsilon^\intercal\Psi^\intercal w + U_Y = \varepsilon^\intercal\Psi^\intercal w + U_Y. \end{aligned}$$

Furthermore, $U_Y = \alpha_{0,-*}^\intercal Z_{-*} + \eta_0^\intercal H + \varepsilon_Y$ can be written as a linear function of A plus a linear function of ε. To realize this, simply express Z_{-*} and H by their marginal reduced form structural equations. Hence, the assumptions that $A \perp\!\!\!\perp U_Y$ must entail that A vanishes from the expression of U_Y. As a consequence we have that $R(\alpha)$ is a linear function only of ε, from which the assumption that $A \perp\!\!\!\perp \varepsilon$ yields that $A \perp\!\!\!\perp R(\alpha)$. That is, the null hypothesis of zero correlation implies independence in the linear structural equation model, under the given assumptions. Thus, $E\|AR(\alpha)\|_2^2 = E\|A\|_2^2 E\|R(\alpha)\|_2^2 < \infty$, so the covariance matrix of $AR(\alpha)$ is well-defined.

By the established independence and Equation (A.12), the covariance matrix of $AR(\alpha)$ has the following representation

$$\text{Cov}(AR(\alpha)) = E(AA^\intercal)E(R(\alpha)^2) \succ 0.$$

The positive definiteness follows from the facts that $E(AA^\intercal) \succ 0$ and $E(R(\alpha)^2) > 0$ for any $\alpha \in \mathbb{R}^{d_1+q_1}$. $E(AA^\intercal) \succ 0$ follows by Assumption 2.1.(h) and $E(R(\alpha)^2) > 0$

for any $\alpha \in \mathbb{R}^{d_1+q_1}$ follows by Assumption 2.1.(b), Assumption 2.1.(c) and Assumption 2.5; non-degeneracy and mutual independence of the marginal noise variables in ε. To see this, expand $R(\alpha)$ in terms of the marginal reduced form structural equations of Y and Z and use that $(I - B^\intercal)$ is invertible to see that ε does not vanish in the expression $R(\alpha)$. The multi-dimensional Central Limit Theorem yields that

$$\frac{1}{\sqrt{n}}\mathbf{A}^\intercal\mathbf{R}(\alpha) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} A_{i,1}R(\alpha)_i \\ \vdots \\ A_{i,q}R(\alpha)_i \end{pmatrix}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Cov}(AR(\alpha))).$$

Furthermore, note that regardless of whether or not the null-hypothesis is true, we have that

$$\sqrt{n}(\mathbf{A}^\intercal\mathbf{A})^{-1/2} \xrightarrow{P} E(AA^\intercal)^{-1/2},$$

and

$$\frac{1}{\sqrt{n}}\|\mathbf{R}(\alpha)\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} R(\alpha)_i^2} \xrightarrow{P} \sqrt{E(R(\alpha)^2)} > 0,$$

by the law of large numbers and the continuity of the matrix square root operation on the cone of symmetric positive-definite matrices. We can represent the test-statistic as $T_n(\alpha) := \|\sqrt{n}W_n(\alpha)\|_2^2$ with

$$W_n(\alpha) := (\mathbf{A}^\intercal\mathbf{A})^{-1/2}\mathbf{A}^\intercal\mathbf{R}(\alpha)/\|\mathbf{R}(\alpha)\|_2,$$

and have that $\sqrt{n}W_n(\alpha) \xrightarrow{\mathcal{D}} W \sim \mathcal{N}(0, I)$, by Slutsky's theorem and linear transformation rules of multivariate normal distributions. Hence, the continuous mapping theorem yields that

$$T_n(\alpha) = \left\|\sqrt{n}W_n(\alpha)\right\|_2^2 \xrightarrow{\mathcal{D}} \|W\|_2^2 = \sum_{i=1}^{q} W_i^2 \sim \chi_q^2,$$

where χ_q^2 is the Chi-squared distribution with q degrees of freedom, since $W_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp W_q$. This proves that the test-statistic T_n has the correct asymptotic distribution under the null-hypothesis.

Now fix a distribution P, for which the null hypothesis of simultaneous zero correlation between the components of A and the residuals $R(\alpha)$ does not hold. That is, there exists an $j \in \{1, ..., q\}$ such that $E(A_jR(\alpha)) \neq E(A_j)E(R(\alpha)) = 0$. Note that

$$\begin{aligned}
&\left\|n^{-1/2}(\mathbf{A}^\intercal\mathbf{A})^{1/2}\right\|_{\mathrm{op}}^2 T_n(\alpha) \\
&= \left\|n^{-1/2}(\mathbf{A}^\intercal\mathbf{A})^{1/2}\right\|_{\mathrm{op}}^2 \left\|\frac{\sqrt{n}(\mathbf{A}^\intercal\mathbf{A})^{-1/2}\frac{1}{\sqrt{n}}\mathbf{A}^\intercal\mathbf{R}(\alpha)}{\frac{1}{\sqrt{n}}\|\mathbf{R}(\alpha)\|_2}\right\|_2^2 \\
&\geq \left\|\frac{\frac{1}{\sqrt{n}}\mathbf{A}^\intercal\mathbf{R}(\alpha)}{\frac{1}{\sqrt{n}}\|\mathbf{R}(\alpha)\|_2}\right\|_2^2 \geq \left|\frac{\frac{1}{\sqrt{n}}\mathbf{A}_j^\intercal\mathbf{R}(\alpha)}{\frac{1}{\sqrt{n}}\|\mathbf{R}(\alpha)\|_2}\right|^2 = n\left|\frac{\frac{1}{n}\mathbf{A}_j^\intercal\mathbf{R}(\alpha)}{\frac{1}{\sqrt{n}}\|\mathbf{R}(\alpha)\|_2}\right|^2,
\end{aligned}$$

where $\mathbf{A}_j^\intercal := (\mathbf{A}_j)^\intercal$ and $\mathbf{A}_j$ is the j'th column of $\mathbf{A}$ corresponding to the i.i.d. vector consisting of n copies of the j'th exogenous variable A_j and $\|\cdot\|_{\text{op}}$ is the operator norm. The lower bound diverges to infinity in probability as the latter factor tends to $|E(A_j R(\alpha))/\sqrt{E(R(\alpha)^2)}|^2 > 0$ in probability by the law of large numbers and Slutsky's theorem. Hence, it holds that $T_n(\alpha) \xrightarrow{P} \infty$, as

$$\left\|n^{-1/2}(\mathbf{A}^\intercal\mathbf{A})^{1/2}\right\|_{\text{op}} \to \left\|E(AA^\intercal)^{1/2}\right\|_{\text{op}} \in (0,\infty).$$

This concludes the proof. □

Proof of Lemma 2.2: Let Assumption 2.6 hold, i.e., that $\mathbf{Z}^\intercal\mathbf{Z}$ and $\mathbf{A}^\intercal\mathbf{Z}$ are of full rank. That $\alpha \mapsto l_{\text{IV}}^n(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A})$ is a convex function and $\alpha \mapsto l_{\text{OLS}}^n(\alpha)$ is a strictly convex function can be seen from the quadratic forms of their second derivatives, i.e.,

$$y^\intercal D^2 l_{\text{IV}}^n(\alpha) y = 2n^{-1} y^\intercal \mathbf{Z}^\intercal\mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{Z}y = 2n^{-1}\|(\mathbf{A}^\intercal\mathbf{A})^{-1/2}\mathbf{A}^\intercal\mathbf{Z}y\|_2^2 \geq 0,$$

and

$$y^\intercal D^2 l_{\text{OLS}}^n(\alpha) y = 2n^{-1} y^\intercal \mathbf{Z}^\intercal\mathbf{Z}y = 2n^{-1}\|\mathbf{Z}y\|_2^2 > 0,$$

for any $y \in \mathbb{R}^{d_1+q_1} \setminus \{0\}$. Here, we also used that $\mathbf{A}^\intercal\mathbf{A}$ is of full rank by Assumption 2.1.(i) and that $\mathbf{Z} \in \mathbb{R}^{n\times(d_1+q_1)}$ is an injective linear transformation as $d_1 + q_1 = \text{rank}(\mathbf{Z}^\intercal\mathbf{Z}) = \text{rank}(\mathbf{Z})$.

Suppose that there exists two optimal solutions α_1, α_2 to the (Primal.$t.n$) problem. By the convexity of the feasibility set any convex combination is also feasible. However,

$$l_{\text{OLS}}^n(\alpha_1/2 + \alpha_2/2) < l_{\text{OLS}}^n(\alpha_1)/2 + l_{\text{OLS}}^n(\alpha_2)/2 = l_{\text{OLS}}^n(\alpha_1),$$

since $l_{\text{OLS}}^n(\alpha_1) = l_{\text{OLS}}^n(\alpha_2)$. This means that $\alpha_1/2 + \alpha_2/2$ has a strictly better objective value than the optimal point α_1, which is a contradiction. Hence, there cannot exist multiple solutions to the optimization problem (Primal.$t.n$).

Regarding the claim of solvability, note that $\mathbf{Z}^\intercal\mathbf{Z}$ is positive definite and as a consequence the smallest eigenvalue $\lambda_{\min}(\mathbf{Z}^\intercal\mathbf{Z})$ is strictly positive. Thus, using the lower bound of the symmetric quadratic form $\alpha^\intercal\mathbf{Z}^\intercal\mathbf{Z}\alpha \geq \lambda_{\min}(\mathbf{Z}^\intercal\mathbf{Z})\|\alpha\|_2^2$, we get that

$$\begin{aligned} l_{\text{OLS}}^n(\alpha) &= \mathbf{Y}^\intercal\mathbf{Y} + \alpha^\intercal\mathbf{Z}^\intercal\mathbf{Z}\alpha - 2\mathbf{Y}^\intercal\mathbf{Z}\alpha \geq \mathbf{Y}^\intercal\mathbf{Y} + \lambda_{\min}(\mathbf{Z}^\intercal\mathbf{Z})\|\alpha\|_2^2 - 2|\mathbf{Y}^\intercal\mathbf{Z}\alpha| \\ &\geq \mathbf{Y}^\intercal\mathbf{Y} + \lambda_{\min}(\mathbf{Z}^\intercal\mathbf{Z})\|\alpha\|_2^2 - 2\|\mathbf{Y}^\intercal\mathbf{Z}\|_{\text{op}}\|\alpha\|_2 \to \infty, \end{aligned} \tag{A.13}$$

as $\|\alpha\|_2 \to \infty$, where we used that for the linear operator $\mathbf{Y}^\intercal\mathbf{Z} : \mathbb{R}^{d_1+q_1} \to \mathbb{R}$ the operator norm is given by $\|\mathbf{Y}^\intercal\mathbf{Z}\|_{\text{op}} := \inf\{c \geq 0 : |\mathbf{Y}\mathbf{Z}v| \leq c\|v\|_2, \forall v \in \mathbb{R}^{d_1+q_1}\}$, obviously satisfying $|\mathbf{Y}^\intercal\mathbf{Z}v| \leq \|\mathbf{Y}^\intercal\mathbf{Z}\|_{\text{op}}\|v\|_2$ for any $v \in \mathbb{R}^{d_1+q_1}$.

Now assume that $t > \inf_\alpha l_{\text{IV}}^n(\alpha)$. This implies that there exists at least one point $\tilde{\alpha} \in \mathbb{R}^{d_1+q_1}$ such that $l_{\text{IV}}^n(\tilde{\alpha}) \leq t$, hence we only need to consider points α

such that $l^n_{\text{OLS}}(\alpha) \leq l^n_{\text{OLS}}(\tilde{\alpha})$ as possible solutions of the optimization problem. By the considerations in Equation (A.13) above, there exists $c \geq 0$ such that is suffices to search over the closed ball $\overline{B(0,c)}$. Indeed, for a sufficiently large $c \geq 0$ we know that $\alpha \notin \overline{B(0,c)}$ implies that $l^n_{\text{OLS}}(\alpha) > l^n_{\text{OLS}}(\tilde{\alpha})$ by Equation (A.13). Furthermore, as the inequality constraint function $\alpha \mapsto l^n_{\text{IV}}(\alpha)$ is continuous, the set of feasible points $(l^n_{\text{IV}})^{-1}((-\infty, t])$ is closed. Hence, our minimization problem is equivalent with the minimization of the continuous function $\alpha \mapsto l^n_{\text{OLS}}(\alpha)$ over the convex and compact set $\overline{B(0,c)} \cap (l^n_{\text{IV}})^{-1}((-\infty, t])$. By the extreme value theorem, the minimum exist and is attainable. We conclude that the primal problem is solvable if $t > \inf_\alpha l^n_{\text{IV}}(\alpha)$.

By definition, Slater's condition is satisfied if there exists a point in the relative interior of the problem domain where the constraint inequality is strict (Boyd and Vandenberghe, 2004). Since the problem domain is $\mathbb{R}^{d_1+q_1}$, we need the existence of $\alpha \in \mathbb{R}^{d_1+q_1}$ such that $l^n_{\text{IV}}(\alpha) < t$. This is clearly satisfied if $t > \inf_\alpha l^n_{\text{IV}}(\alpha)$. Let us now specify the exact lower bound for the constraint bound as a function of the over-identifying restrictions. *Under- and just-identified case: $q_2 \leq d_1$ ($q \leq d_1 + q_1$).* Assumption 2.6.(b) yields that $\mathbf{A}^\intercal\mathbf{Z} \in \mathbb{R}^{q\times(d_1+q_1)}$ satisfies $\text{rank}(\mathbf{A}^\intercal\mathbf{Z}) = q$. That is, $\mathbf{A}^\intercal\mathbf{Z}$ is of full row rank, hence surjective. Thus, we are guaranteed the existence of a $\tilde{\alpha} \in \mathbb{R}^{d_1+q_1}$ such that $\mathbf{A}^\intercal\mathbf{Z}\tilde{\alpha} = \mathbf{A}^\intercal\mathbf{Y}$, implying that $l^n_{\text{IV}}(\tilde{\alpha}) = 0$. *Over-identified case: $d_1 < q_2$ ($d_1 + q_1 < q$).* Note that the constraint function $l^n_{\text{IV}}(\alpha) : \mathbb{R}^{d_1+q_1} \to \mathbb{R}$ is strictly convex as the second derivative $D^2 l^n_{\text{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A}) \propto \mathbf{Z}^\intercal\mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{Z}$ is positive definite by the assumption that $\mathbf{A}^\intercal\mathbf{Z} \in \mathbb{R}^{q\times(d_1+q_1)}$ has full (column) rank. The global minimum of l_{IV} is therefore attained in the unique stationary point. Furthermore, the stationary point is found by solving the normal equation $Dl^n_{\text{IV}}(\alpha; \mathbf{Y}, \mathbf{Z}, \mathbf{A}) = 0$. The solution to the normal equation is given by $\hat{\alpha}^n_{\text{TSLS}} = (\mathbf{Z}^\intercal P_{\mathbf{A}}\mathbf{Z})^\intercal\mathbf{Z}^\intercal P_{\mathbf{A}}\mathbf{Y}$, which is the standard TSLS estimator. □

Proof of Theorem 2.2: Let $p_{\min} \in (0, 1)$ and let Assumption 2.6 and Assumption 2.7 hold. That is, $\mathbf{A}^\intercal\mathbf{Z}$ and $\mathbf{Z}^\intercal\mathbf{Z}$ are of full rank and $[\mathbf{Z}\ \mathbf{Y}]$ is of full column rank. Furthermore, assume that $t^\star_n(p_{\min}) > -\infty$ and $T_n(\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p_{\min}))) \leq Q_{\chi^2_q}(1 - p_{\min})$. First assume that $\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p_{\min})) = \hat{\alpha}^n_{\text{OLS}}$. We note that

$$T_n(\hat{\alpha}^n_{\text{OLS}}) = T_n(\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p_{\min}))) \leq Q_{\chi^2_q}(1 - p_{\min}),$$

hence the global minimizer $\hat{\alpha}^n_{\text{OLS}}$ of $\alpha \mapsto l^n_{\text{OLS}}(\alpha)$ is unique, feasible and necessarily optimal in the PULSE problem, so $\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p_{\min})) = \hat{\alpha}^n_{\text{OLS}} = \hat{\alpha}^n_{\text{PULSE}}$ and we are done.

Now assume that $\hat{\alpha}^n_{\text{Pr}}(t^\star_n(p)) \neq \hat{\alpha}^n_{\text{OLS}}$. Consider the PULSE problem of interest

$$\begin{array}{ll} \min_\alpha & l^n_{\text{OLS}}(\alpha) \\ \text{subject to} & T_n(\alpha) \leq Q_{\chi^2_q}(1 - p_{\min}), \end{array} \tag{PULSE}$$

which is, in general, a non-convex quadratically constrained quadratic program. First we argue that the problem is solvable, i.e., the optimum is attainable.

> To see this, let $p = p_{\min}$, $Q = Q_{\chi^2_q}(1 - p_{\min})$ and note that by the assumption $t^\star_n(p_{\min}) > -\infty$ we have that the feasible set of the PULSE

problem is non-empty. By the assumptions that $[\mathbf{Z}\ \mathbf{Y}]$ is of full column rank we have that $T_n(\alpha)$ is well-defined for any $\alpha \in \mathbb{R}^{d_1+q_1}$, as the denominator is never zero. By continuity of $\mathbb{R}^{d_1+q_1} \ni \alpha \mapsto T_n(\alpha)$ we have that the feasible set $\mathcal{F} := T_n^{-1}((-\infty, Q])$, is closed and non-empty, since it is the continuous preimage of a closed set. Applying the same arguments as seen earlier in the proof of Lemma 2.2, we know that $l^n_{\mathrm{OLS}}(\alpha) \to \infty$ when $\|\alpha\| \to \infty$. Hence, for a sufficiently large $c > 0$ we know that if $\alpha \notin \overline{B(0,c)}$, where $\overline{B(0,c)} \subseteq \mathbb{R}^{d_1+q_1}$ is the closed ball with centre 0 and radius c, then we only get suboptimal objective values $l^n_{\mathrm{OLS}}(\alpha) > l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p)))$. That is, we can without loss of optimality or loss of solutions restrict the feasible set to $\mathcal{F}' = T_n^{-1}((-\infty, Q]) \cap \overline{B(0,c)}$ a closed and bounded set in $\mathbb{R}^{d_1+q_1}$. Hence, by the extreme value theorem the minimum over $\mathcal{F}'$ is guaranteed to be attained. That is, the PULSE problem is solvable.

However, by the non-convexity of T_n, the preimage $T_n^{-1}((-\infty, Q])$ is in general not convex, so the minimum is not yet guaranteed to be attained in a unique point. We will show that the minimum of the PULSE problem is attained in a unique point, that exactly coincides with the primal PULSE solution. Fix any solution $\hat{\alpha}$ to the PULSE problem and realize that the PULSE constraint is active in $\hat{\alpha}$,

$$T_n(\hat{\alpha}) = Q. \tag{A.14}$$

This is seen by noting that $\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p)) \neq \hat{\alpha}^n_{\mathrm{OLS}}$ by assumption, so $\hat{\alpha}^n_{\mathrm{OLS}}$ is not feasible in the PULSE problem, that is, $\hat{\alpha}^n_{\mathrm{OLS}} \notin \mathcal{F}$. If $\hat{\alpha}^n_{\mathrm{OLS}}$ was feasible, then $t^\star_n(p) = \sup\{t \in D_{\mathrm{Pr}} : T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t)) \leq Q_{\chi^2_q}(1 - p_{\min})\} = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})$, since

$$T_n(\hat{\alpha}^n_{\mathrm{Pr}}(l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})) = T_n(\hat{\alpha}^n_{\mathrm{OLS}}) \leq Q,$$

hence

$$\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p)) = \underset{\alpha : l^n_{\mathrm{IV}}(\alpha) \leq l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})}{\arg\min}\ l^n_{\mathrm{OLS}}(\alpha) = \hat{\alpha}^n_{\mathrm{OLS}},$$

which is a contradiction. That the optimum must be attained in a point, where the PULSE inequality constraint is active then follows from Lemma A.4 of Appendix A.6 and the conclusion above that the only stationary point of l^n_{OLS}, $\hat{\alpha}^n_{\mathrm{OLS}}$, is not feasible.

Thus,

$$T_n(\hat{\alpha}) = n\frac{l^n_{\mathrm{IV}}(\hat{\alpha})}{l^n_{\mathrm{OLS}}(\hat{\alpha})} = Q \iff l^n_{\mathrm{IV}}(\hat{\alpha}) = \frac{Q}{n} l^n_{\mathrm{OLS}}(\hat{\alpha}). \tag{A.15}$$

Furthermore, the assumption that $T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p))) \leq Q$ means that the solution to the primal PULSE, $\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p))$, is feasible in the PULSE problem. That is, $\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p)) \in \mathcal{F}$. As a consequence of this we have that

$$l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p))) \geq \min_{\alpha \in \mathcal{F}} l^n_{\mathrm{OLS}}(\alpha) = l^n_{\mathrm{OLS}}(\hat{\alpha}). \tag{A.16}$$

Now we show that the PULSE solution $\hat{\alpha}$ is feasible in the primal PULSE problem (Primal.$t_n^*(p).n$).

To see this, Note that the feasibility set of the PULSE problem can be shrunk in the following manner

$$\begin{aligned}
\mathcal{F} &= \left\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\mathrm{IV}}^n(\alpha) \leq \frac{Q}{n} l_{\mathrm{OLS}}^n(\alpha)\right\} \\
&= \left\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\mathrm{IV}}^n(\alpha) \leq \frac{Q}{n} l_{\mathrm{OLS}}^n(\alpha), l_{\mathrm{OLS}}^n(\alpha) \geq l_{\mathrm{OLS}}^n(\hat{\alpha})\right\} \\
&\supseteq \left\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\mathrm{IV}}^n(\alpha) \leq \frac{Q}{n} l_{\mathrm{OLS}}^n(\hat{\alpha}), l_{\mathrm{OLS}}^n(\alpha) \geq l_{\mathrm{OLS}}^n(\hat{\alpha})\right\} \\
&= \left\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\mathrm{IV}}^n(\alpha) \leq l_{\mathrm{IV}}^n(\hat{\alpha}), l_{\mathrm{OLS}}^n(\alpha) \geq l_{\mathrm{OLS}}^n(\hat{\alpha})\right\} \\
&= \left\{\alpha \in \mathbb{R}^{d_1+q_1} : l_{\mathrm{IV}}^n(\alpha) \leq l_{\mathrm{IV}}^n(\hat{\alpha})\right\} =: \hat{\mathcal{F}}(\hat{\alpha}),
\end{aligned}$$

where the third equality follows from Equation (A.15). The only claim above that needs justification is that:

$$l_{\mathrm{IV}}^n(\alpha) \leq l_{\mathrm{IV}}^n(\hat{\alpha}) \implies l_{\mathrm{OLS}}^n(\alpha) \geq l_{\mathrm{OLS}}^n(\hat{\alpha}). \tag{A.17}$$

For now we assume that this claim holds and provide a proof later. Thus, we have that $\hat{\mathcal{F}}(\hat{\alpha}) \subseteq \mathcal{F}$ and we note that $\hat{\alpha} \in \hat{\mathcal{F}}(\hat{\alpha})$. An important consequence of this is that the PULSE solution $\hat{\alpha}$ is also the unique solution to the primal problem (Primal.$l_{\mathrm{IV}}^n(\hat{\alpha}).n$). That is,

$$\hat{\alpha} = \hat{\alpha}_{\mathrm{Pr}}^n(l_{\mathrm{IV}}^n(\hat{\alpha})) = \begin{array}{ll} \operatorname{argmin}_\alpha & l_{\mathrm{OLS}}^n(\alpha) \\ \text{subject to} & l_{\mathrm{IV}}^n(\alpha) \leq l_{\mathrm{IV}}^n(\hat{\alpha}). \end{array}$$

We will now prove that $l_{\mathrm{IV}}^n(\hat{\alpha}) \in \mathcal{E} := \{t \in [\min_\alpha l_{\mathrm{IV}}^n(\alpha), l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{OLS}}^n)] : T_n(\hat{\alpha}_{\mathrm{Pr}}^n(t)) \leq Q_{\chi_q^2}(1-p)\}$. This follows from the following two observations: (1) $\min_\alpha l_{\mathrm{IV}}^n(\alpha) \leq l_{\mathrm{IV}}^n(\hat{\alpha}) < l_{\mathrm{IV}}^n(\hat{\alpha}_{\mathrm{OLS}}^n)$ and (2) $T_n(\hat{\alpha}_{\mathrm{Pr}}^n(l_{\mathrm{IV}}^n(\hat{\alpha}))) \leq Q_{\chi_q^2}(1-p)$. (1) follows because $\hat{\alpha}_{\mathrm{OLS}}^n \notin \mathcal{F}$, which implies, by the above inclusion, that $\hat{\alpha}_{\mathrm{OLS}}^n \notin \hat{\mathcal{F}}(\hat{\alpha})$. (2) follows because $\hat{\alpha}$ solves (Primal.$l_{\mathrm{IV}}^n(\hat{\alpha}).n$) and thus $\hat{\alpha}_{\mathrm{Pr}}^n(l_{\mathrm{IV}}^n(\hat{\alpha})) = \hat{\alpha}$; $T_n(\hat{\alpha}) \leq Q_{\chi_q^2}(1-p)$ holds because $\hat{\alpha}$ is feasible for the PULSE problem.

Now, since $t_n^\star(p) = \sup(\mathcal{E} \setminus \{\min_\alpha l_{\mathrm{IV}}^n(\alpha)\}) \in \mathbb{R}$ implies $t_n^\star(p) = \sup(\mathcal{E})$, it follows that $l_{\mathrm{IV}}^n(\hat{\alpha}) \leq t_n^\star(p)$. In other words, any solution $\hat{\alpha}$ to the PULSE problem is feasible in the primal PULSE problem (Primal.$t_n^*(p).n$).

Hence,

$$l_{\mathrm{OLS}}^n(\hat{\alpha}) \geq l_{\mathrm{OLS}}(\hat{\alpha}_{\mathrm{Pr}}^n(t_n^\star(p))). \tag{A.18}$$

Equation (A.16) and Equation (A.18) now yield that $l_{\mathrm{OLS}}^n(\hat{\alpha}_{\mathrm{Pr}}^n(t_n^\star(p))) = l_{\mathrm{OLS}}^n(\hat{\alpha})$ for any PULSE solution $\hat{\alpha}$. Thus, any solution $\hat{\alpha}$ to the PULSE problem is feasible

in the primal PULSE problem (Primal.$t_n^*(p).n$) and it attains the optimal primal PULSE objective value. We conclude that $\hat{\alpha}$ solves the primal PULSE problem. Furthermore, it must hold that $\hat{\alpha} = \hat{\alpha}^n_{\mathrm{Pr}}(t_n^\star(p))$, by uniqueness of solutions to the primal PULSE problem(see Lemma 2.2). This implies two things: solutions to the PULSE problem are unique and the PULSE solution coincides with the primal PULSE solution.

It only remains to prove the claim of Equation (A.17), which ensures $\hat{\mathcal{F}}(\hat{\alpha}) \subseteq \mathcal{F}$. Assume for contradiction that there exists an α such that $l^n_{\mathrm{IV}}(\alpha) \leq l^n_{\mathrm{IV}}(\hat{\alpha})$ and $l^n_{\mathrm{OLS}}(\alpha) < l^n_{\mathrm{OLS}}(\hat{\alpha})$, that is, we assume that

$$\mathcal{A} := \underbrace{\{\alpha \in \mathbb{R}^{d_1+q_1} : l^n_{\mathrm{OLS}}(\alpha) < l^n_{\mathrm{OLS}}(\hat{\alpha})\}}_{=:\mathcal{B}} \cap \underbrace{\{\alpha \in \mathbb{R}^{d_1+q_1} : l^n_{\mathrm{IV}}(\alpha) \leq l^n_{\mathrm{IV}}(\hat{\alpha})\}}_{=:\mathcal{C}} \neq \emptyset.$$

Define $\mathcal{M}_{\mathrm{IV}} := \{\alpha : l^n_{\mathrm{IV}}(\alpha) = \min_{\alpha'} l^n_{\mathrm{IV}}(\alpha')\}$ as the solution space to the generalized method of moments formulation of the instrumental variable minimization problem. We now prove that $\mathcal{M}_{\mathrm{IV}} \cap \mathcal{A} = \emptyset$.

> That is, we claim that in the just- and over-identified setup $\hat{\alpha}^n_{\mathrm{TSLS}} \notin \mathcal{A}$ and in the under-identified setup none of the infinitely many solutions in the solution space of the instrumental variable minimization problem lies in $\mathcal{A}$. These statements follow by first noting that $\mathcal{M}_{\mathrm{IV}} \subseteq \mathcal{F}$ in any identification setting. In the under- and -just identified setup this is seen by noting that $l^n_{\mathrm{IV}}(\alpha) = 0$ for any $\alpha \in \mathcal{M}_{\mathrm{IV}}$, which implies $T_n(\alpha) = 0 \leq Q$, hence $\mathcal{M}_{\mathrm{IV}} \subseteq \mathcal{F}$. In the over-identified setup, where $\mathcal{M}_{\mathrm{IV}} = \{\hat{\alpha}^n_{\mathrm{TSLS}}\}$, we will now argue that $\mathcal{M}_{\mathrm{IV}} \subseteq \mathcal{F}$ follows from the assumption that $t_n^\star(p) < \infty$. We first prove that $D_{\mathrm{Pr}} \ni t \mapsto T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t))$ is weakly increasing. If $t_1 < t_2$ are two constraint bounds for which the primal problem is solvable, then $l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1)) \geq l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2))$ as the feasibility set for t_2 is larger than the one for t_1. Furthermore, the solution $\hat{\alpha}^n_{\mathrm{Pr}}(t_2)$ either equals $\hat{\alpha}^n_{\mathrm{Pr}}(t_1)$ or is contained in the set $\{\alpha \in \mathbb{R}^{d_1+q_1} : t_1 < l^n_{\mathrm{IV}}(\alpha) \leq t_2\}$; in the latter case we have $l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1)) \leq t_1 < l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2)) \leq t_2$. Thus, we have in both cases that $l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1)) \leq l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2))$. Combining the two observations above we have that
>
> $$T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t_1)) = n\frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1))} \leq n\frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2))} = T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t_2)).$$
>
> Hence, as $-\infty < \min_\alpha l^n_{\mathrm{IV}}(\alpha) = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{TSLS}}) < t_n^\star(p) < \infty$ are two points for which the primal problem is solvable we get that
>
> $$T_n(\hat{\alpha}^n_{\mathrm{TSLS}}) = T_n(\hat{\alpha}^n_{\mathrm{Pr}}(l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{TSLS}}))) \leq T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t_n^\star(p))) \leq Q.$$
>
> This proves that $\mathcal{M}_{\mathrm{IV}} \subseteq \mathcal{F}$ in the over-identified setup. Now, if $\mathcal{M}_{\mathrm{IV}} \cap \mathcal{A} \neq \emptyset$, there exists an $\alpha \in \mathcal{M}_{\mathrm{IV}} \cap \mathcal{A} \subseteq \mathcal{F} \cap \mathcal{A}$ such that α is feasible in the PULSE problem ($\alpha \in \mathcal{F}$) and α is super-optimal compared to $\hat{\alpha}$, $l^n_{\mathrm{OLS}}(\alpha) < l^n_{\mathrm{IV}}(\hat{\alpha})$ ($\alpha \in \mathcal{A}$), contradicting that $\hat{\alpha}$ is an solution to the PULSE problem. We can thus conclude that $\mathcal{M}_{\mathrm{IV}} \cap \mathcal{A} = \emptyset$.

This allows us to fix two distinct points $\bar{\alpha} \neq \alpha'$ such that $\bar{\alpha} \in \mathcal{A}$ and $\alpha' \in \mathcal{M}_{\text{IV}}$. Consider the proper line segment function between $\bar{\alpha}$ and α', $f(t) : [0,1] \to \mathbb{R}^{d_1+q_1}$ given by $f(t) := t\alpha' + (1-t)\bar{\alpha}$. A multivariate convex function is convex in any direction from any given starting point in its domain, so both $l^n_{\text{IV}} \circ f : [0,1] \to \mathbb{R}_+$ and $l^n_{\text{OLS}} \circ f : [0,1] \to \mathbb{R}_+$ are convex. Since $\mathcal{M}_{\text{IV}} \cap \mathcal{A} = \emptyset$ it is obvious that the function f will for sufficiently large t 'leave' the set $\mathcal{A}$. We will now prove that f actually leaves the superset $\mathcal{B} \supset \mathcal{A}$. More precisely, we will prove that there exists a $t_1 \in (0,1]$ such that for all $t' \in [0, t_1)$ it holds that $f(t') \in \mathcal{B}$ and for all $t' \in [t_1, 1]$ it holds that $f(t') \notin \mathcal{B}$ (which implies $f(t') \notin \mathcal{A}$).

> Because $l^n_{\text{IV}}(\alpha') = \min_\alpha l^n_{\text{IV}}(\alpha)$ we have that $\alpha' \in \mathcal{C} = \{\alpha : l^n_{\text{IV}}(\alpha) \leq l^n_{\text{IV}}(\hat{\alpha})\}$. By convexity of l^n_{IV} (Lemma 2.2) the sublevel set $\mathcal{C}$ is convex and thus contains the entire line segment between $\bar{\alpha}$ and α'. As a consequence $a' \notin \mathcal{B}$. It therefore suffices to construct a $t_1 \in (0,1]$ such that for all $t' \in [0, t_1)$ it holds that $f(t') \in \mathcal{B}$ and for all $t' \in [t_1, 1]$ it holds that $f(t') \notin \mathcal{B}$. We now consider the set $\{t \in [0,1] : l^n_{\text{OLS}}(f(t)) < l^n_{\text{OLS}}(\hat{\alpha})\} = f^{-1}(\mathcal{B})$. This set contains 0 because $\bar{\alpha} \in \mathcal{A} \subseteq \mathcal{B}$; it does not contain 1 because $\alpha' \notin \mathcal{B}$; it is convex, as it is a sublevel set of a convex function ($l^n_{\text{OLS}} \circ f$); it is relatively open in $[0,1]$ because it is a pre-image of an open set under a continuous function ($l^n_{\text{OLS}} \circ f$). Thus, the set must be of the form $[0, t_1)$ for some $t_1 \in (0,1]$. This t_1 satisfies the desired criteria.

We constructed t_1 above such that for all $t' \in [0, t_1)$ it holds that $l^n_{\text{OLS}}(f(t')) < l^n_{\text{OLS}}(\hat{\alpha})$ and for all $t' \in [t_1, 1]$ it holds that $l^n_{\text{OLS}}(f(t')) \geq l^n_{\text{OLS}}(\hat{\alpha})$. By continuity of $l^n_{\text{OLS}} \circ f$ we must therefore have that $l^n_{\text{OLS}}(f(t_1)) = l^n_{\text{OLS}}(\hat{\alpha})$. Since $f(1) = \alpha'$ is a global minimum for l^n_{IV}, we have that 1 must also be a global minimum for $l^n_{\text{IV}} \circ f$, implying that the convex the function $l^n_{\text{IV}} \circ f : [0,1] \to \mathbb{R}_+$ is monotonically decreasing. It must therefore hold that

$$l^n_{\text{IV}}(f(t_1)) < l^n_{\text{IV}}(f(0)) = l^n_{\text{IV}}(\bar{a}) \leq l^n_{\text{IV}}(\hat{\alpha}).$$

The first inequality is strict because if $l^n_{\text{IV}}(f(t_1)) = l^n_{\text{IV}}(f(0)) = l^n_{\text{IV}}(\bar{\alpha})$, then convexity of l^n_{IV} implies that

$$l^n_{\text{IV}}(f(t_1)) = l^n_{\text{IV}}(t_1\alpha' + (1-t_1)\bar{\alpha}) \leq t_1 l^n_{\text{IV}}(\alpha') + (1-t_1) l^n_{\text{IV}}(\bar{\alpha}),$$

which happens if and only if $l^n_{\text{IV}}(\bar{\alpha}) \leq l^n_{\text{IV}}(\alpha')$ contradicting the already established fact that $l^n_{\text{IV}}(\bar{\alpha}) > l^n_{\text{IV}}(\alpha')$, which holds since $\alpha' \in \mathcal{M}_{\text{IV}}$ but $\bar{\alpha} \notin \mathcal{M}_{\text{IV}}$. We conclude that $l^n_{\text{IV}}(f(t_1)) < l^n_{\text{IV}}(\hat{\alpha})$.

Thus, we have argued that $\mathcal{M}_{\text{IV}} \cap \mathcal{A} = \emptyset$ implies the existence of an $\tilde{\alpha} := f(t_1) = t_1\alpha' + (1-t_1)\bar{\alpha}$ such that $l^n_{\text{IV}}(\tilde{\alpha}) < l^n_{\text{IV}}(\hat{\alpha})$ and $l^n_{\text{OLS}}(\tilde{\alpha}) = l^n_{\text{OLS}}(\hat{\alpha})$. We have illustrated the above considerations in Figure A.1. It follows that

$$T_n(\tilde{\alpha}) = n\frac{l^n_{\text{IV}}(\tilde{\alpha})}{l^n_{\text{OLS}}(\tilde{\alpha})} = n\frac{l^n_{\text{IV}}(\tilde{\alpha})}{l^n_{\text{OLS}}(\hat{\alpha})} < n\frac{l^n_{\text{IV}}(\hat{\alpha})}{l^n_{\text{OLS}}(\hat{\alpha})} = Q,$$

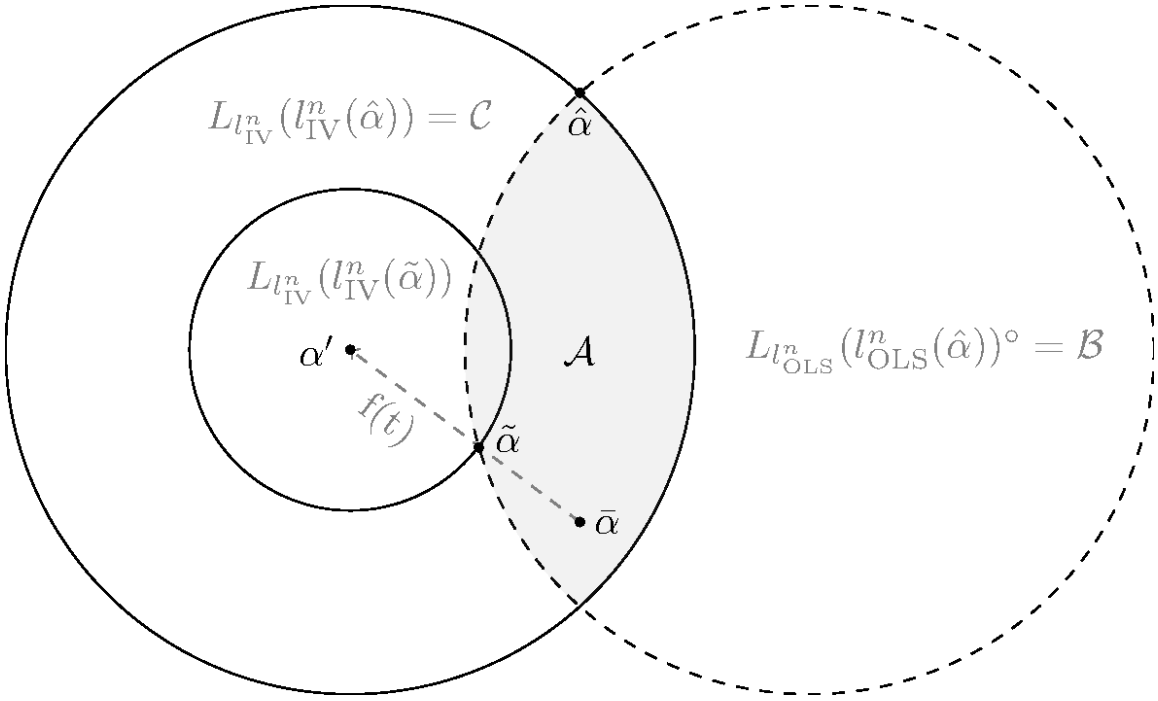


Figure A.1: Illustration of the described procedure in the just- or over-identified setup with $d_1 + q_1 = 2$, where we show that $\mathcal{A} \neq \emptyset$ leads to a contradiction. Here, $L_g(c) := \{\alpha : g(\alpha) \leq c\}$ is the c sublevel set of the function g and A° denotes the interior of a set A. The illustration is simplified, e.g., because the sublevel sets are convex but not necessarily Euclidean balls. Note that the position of $\hat{\alpha}^n_{\mathrm{OLS}}$ is not specified, as it can possibly be in either $\mathcal{A}$ or $L_{l^n_{\mathrm{OLS}}}(l^n_{\mathrm{OLS}}(\hat{\alpha})) \setminus \mathcal{A}$. In the under-indentified setup α' would lie in the $d - q_2 = 1$ dimensional subspace $\mathcal{M}_{\mathrm{IV}}$ and the level sets would be slabs around this line.

implying that $\tilde{\alpha}$ is strictly feasible in the PULSE problem and, in fact, a solution as the objective value is optimal. We argued earlier in Equation (A.14) that any solution to the PULSE problem must be tight in the inequality constraint, hence we have arrived at a contradiction. We conclude that $\mathcal{A} = \emptyset$, which implies that Equation (A.17) must hold. □

Proof of Lemma 2.3: Assume that we are in the just- or over-identified setup and that Assumption 2.6 are satisfied. That is, $\mathbf{Z}^\intercal\mathbf{Z}$, $\mathbf{A}^\intercal\mathbf{Z}$ and $\mathbf{A}^\intercal\mathbf{A}$ are almost surely of full rank. In particular we have that $\mathbf{Z}$, $\mathbf{A}^\intercal\mathbf{Z}$ and $P_{\mathbf{A}}\mathbf{Z}$ are almost surely of full column rank (injective linear maps). Furthermore, let ε have density with respect to the Lebesgue measure and let B be lower triangular. Fix $\lambda \geq 0$ and $\omega \in W_\lambda$, where

$$\begin{aligned} W_\lambda :=& (\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}) \cap (\mathrm{rank}(\mathbf{Z}^\intercal\mathbf{Z}) = d_1 + q_1) \\ & \cap (\mathrm{rank}(\mathbf{A}^\intercal\mathbf{Z}) = d_1 + q_1) \cap (\mathrm{rank}(\mathbf{A}^\intercal\mathbf{A}) = q), \end{aligned}$$

satisfying $P(\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}) = P(W_\lambda)$. By Equation (2.13) we have that

$$\begin{aligned}
\hat{\alpha}^n_{\mathrm{K}}(\lambda) &= \arg\min_\alpha \{l^n_{\mathrm{OLS}}(\alpha) + \lambda l^n_{\mathrm{IV}}(\alpha)\} \\
&= \arg\min_\alpha \{(\mathbf{Y} - \mathbf{Z}\alpha)^\intercal(\mathbf{Y} - \mathbf{Z}\alpha) + \lambda(\mathbf{Y} - \mathbf{Z}\alpha)^\intercal P_{\mathbf{A}}(\mathbf{Y} - \mathbf{Z}\alpha)\} \\
&= \arg\min_\alpha (\mathbf{Y} - \mathbf{Z}\alpha)^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})(\mathbf{Y} - \mathbf{Z}\alpha) \\
&= \arg\min_\alpha \|(\mathbf{I} + \lambda P_{\mathbf{A}})^{1/2}(\mathbf{Y} - \mathbf{Z}\alpha)\|_2^2 \\
&= \arg\min_\alpha \|\mathbf{Y} - \mathbf{Z}\alpha\|^2_{(\mathbf{I}+\lambda P_{\mathbf{A}})},
\end{aligned}$$

where $\|\cdot\|_{(\mathbf{I}+\lambda P_{\mathbf{A}})}$ is the norm induced by the inner product $\langle x, y\rangle_{(\mathbf{I}+\lambda P_{\mathbf{A}})} = x^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})y$. The solution $\mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is well-known to coincide with the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{R}(\mathbf{Z})$, the range of $\mathbf{Z}$, with respect to the inner product $\langle\cdot,\cdot\rangle_{(\mathbf{I}+\lambda P_{\mathbf{A}})}$. Hence, $\mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is the unique element in this closed linear subspace such that for all $z \in \mathcal{R}(\mathbf{Z})$ it holds that

$$\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), z\rangle_{(\mathbf{I}+\lambda P_{\mathbf{A}})} = \langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), (\mathbf{I} + \lambda P_{\mathbf{A}})z\rangle = 0,$$

or equivalently,

$$\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), z\rangle = -\lambda\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), P_{\mathbf{A}}z\rangle, \quad \forall z \in \mathcal{R}(\mathbf{Z}). \tag{A.19}$$

We note that if $\lambda = 0$ then $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{OSL}}$, seen either by directly inspecting the closed form solution of $\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ or concluding the same from Equation (A.19).

Furthermore, when $\lambda > 0$ we have that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}$ implies that, again, $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{OSL}}$. To see this, we note that

$$\hat{\alpha}^n_{\mathrm{TSLS}} = \arg\min_\alpha l^n_{\mathrm{IV}}(\alpha) = \arg\min_\alpha \|P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\alpha\|_2^2,$$

so $P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}$ is the orthogonal projection of $P_{\mathbf{A}}\mathbf{Y}$ onto $\mathcal{R}(P_{\mathbf{A}}\mathbf{Z})$. That is, $P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}$ is the unique element in $\mathcal{R}(P_{\mathbf{A}}\mathbf{Z})$ such that

$$\langle P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}, s\rangle = 0,$$

for all $s \in \mathcal{R}(P_{\mathbf{A}}\mathbf{Z})$, i.e.,

$$\langle P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}, P_{\mathbf{A}}z\rangle = 0,$$

for all $z \in \mathcal{R}(\mathbf{Z})$. Thus, if $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}$ for some $\lambda > 0$ we have that

$$\begin{aligned}
0 &= \langle P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}, P_{\mathbf{A}}z\rangle \\
&= \langle P_{\mathbf{A}}\mathbf{Y} - P_{\mathbf{A}}\mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), P_{\mathbf{A}}z\rangle \\
&= \langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), P_{\mathbf{A}}z\rangle \\
&= -\lambda^{-1}\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), z\rangle,
\end{aligned}$$

hence $\langle \mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{K}}(\lambda), z\rangle = 0$ for all $z \in \mathcal{R}(\mathbf{Z})$, where we used Equation (A.19) and in the third equality we used that $P_{\mathbf{A}}$ is idempotent and symmetric. This implies that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{OLS}}$, as it satisfies the uniquely determining condition for the ordinary least square estimator.

Hence, for any $\lambda \geq 0$, whenever $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}$ we know that $\hat{\alpha}^n_{\mathrm{TSLS}} = \hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{OLS}}$. Thus, for any $\lambda \geq 0$ it holds that

$$P(\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}) \leq P(\hat{\alpha}^n_{\mathrm{TSLS}} = \hat{\alpha}^n_{\mathrm{OLS}}).$$

Recall that the reduced form equations of our system are given by $[\mathbf{Y}\ \mathbf{X}\ \mathbf{H}] = \mathbf{A}\Pi + \boldsymbol{\varepsilon}\Gamma^{-1}$ where $\Gamma := I - B$. When B is lower triangular, so is Γ and Γ^{-1}. By selecting the relevant columns of Π and Γ^{-1} we may express the marginal reduced form structural equations of $\mathbf{S}$ that consist of any collection of columns from $[\mathbf{Y}\ \mathbf{X}\ \mathbf{H}]$ by $\mathbf{S} = \mathbf{A}\Pi_S + \boldsymbol{\varepsilon}\Gamma_S^{-1}$ for conformable matrices Π_S and Γ_S^{-1}. In particular, we have that the marginal reduced form structural equations for $\mathbf{Y}$ and $\mathbf{X}_*$ are given by $\mathbf{Y} = \mathbf{A}\Pi_Y + \boldsymbol{\varepsilon}\Gamma_Y^{-1}$ and $\mathbf{X}_* = \mathbf{A}\Pi_{X_*} + \boldsymbol{\varepsilon}\Gamma_{X_*}^{-1}$, where $\Pi_Y, \Pi_{X_*}, \Gamma_Y^{-1}$ and $\Gamma_{X_*}^{-1}$ are matrices conformable with the following block representation

$$\Pi = [\underbrace{\Pi_Y}_{q\times 1}\ \underbrace{\Pi_{X_*}}_{q\times d_1}\ \underbrace{\Pi_{X_{-*}}}_{d\times q_2}\ \underbrace{\Pi_H}_{q\times r}] \in \mathbb{R}^{q\times l},$$

and

$$\Gamma^{-1} = [\underbrace{\Gamma_Y^{-1}}_{l\times 1}\ \underbrace{\Gamma_{X_*}^{-1}}_{l\times d_1}\ \underbrace{\Gamma_{X_{-*}}^{-1}}_{l\times d_2}\ \underbrace{\Gamma_H^{-1}}_{l\times r}] \in \mathbb{R}^{l\times l},$$

where $l := 1 + d + r$. Note that by the lower triangular structure of Γ^{-1} we have that the only matrix among Γ_Y^{-1}, $\Gamma_{X_*}^{-1}$, $\Gamma_{X_{-*}}^{-1}$ and Γ_H^{-1} that has a non-zero first row is Γ_Y^{-1}.

Now assume without loss of generality that the first row of Γ^{-1} is given by the first canonical Euclidean basis vector $(1, 0, ..., 0) \in \mathbb{R}^{1\times l}$ such that we have the following partitionings

$$\boldsymbol{\varepsilon} = [\underbrace{\boldsymbol{\varepsilon}_Y}_{n\times 1}\ \underbrace{\boldsymbol{\varepsilon}_{-Y}}_{n\times(d+r)}] \in \mathbb{R}^{n\times l}, \qquad \Gamma_Y^{-\intercal} = [1\ \underbrace{\Gamma_{-Y,Y}^{-\intercal}}_{1\times(d+r)}] \in \mathbb{R}^{1\times l},$$

$$\Gamma_{X_*}^{-\intercal} = [\mathbf{0}_{d_1\times 1}\ \underbrace{\Gamma_{-Y,X_*}^{-\intercal}}_{d_1\times(d+r)}] \in \mathbb{R}^{d_1\times l}, \qquad \Gamma_{X_1}^{-\intercal} = [0\ \underbrace{\Gamma_{-Y,X_1}^{-\intercal}}_{1\times(d+r)}] \in \mathbb{R}^{1\times l},$$

where $\mathbf{X}_1$ is the first column of $\mathbf{X}$. Hence, we note that $\boldsymbol{\varepsilon}\Gamma_Y^{-1} = \boldsymbol{\varepsilon}_Y + \boldsymbol{\varepsilon}_{-Y}\Gamma_{-Y,Y}^{-1}$, such that the marginal reduced form structural equation for $\mathbf{Y}$ has the following representation

$$\mathbf{Y} = \mathbf{A}\Pi_Y + \boldsymbol{\varepsilon}\Gamma_Y^{-1} = \mathbf{A}\Pi_Y + \boldsymbol{\varepsilon}_{-Y}\Gamma_{-Y,Y}^{-1} + \boldsymbol{\varepsilon}_Y =: f_y(\mathbf{A}, \boldsymbol{\varepsilon}_{-Y}) + \boldsymbol{\varepsilon}_Y.$$

We can also represent $\mathbf{Z}$ in terms of these structural coefficient block matrices by

$$\begin{aligned}\mathbf{Z} &= \begin{bmatrix}\mathbf{X}_* & \mathbf{A}_*\end{bmatrix} = \begin{bmatrix}\mathbf{A}\Pi_{X_*} + \boldsymbol{\varepsilon}\Gamma_{X_*}^{-1} & \mathbf{A}_*\end{bmatrix} = \begin{bmatrix}\mathbf{A}\Pi_{X_*} & \mathbf{A}_*\end{bmatrix} + \boldsymbol{\varepsilon}\begin{bmatrix}\Gamma_{X_*}^{-1} & \mathbf{0}_{l\times q_1}\end{bmatrix} \\ &= \mathbf{A}\begin{bmatrix}\Pi_{X_*} & \begin{bmatrix}\mathbf{I}_{q_1\times q_1} \\ \mathbf{0}_{q_2\times q_1}\end{bmatrix}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_{-Y}\Gamma_{-Y,X_*}^{-1} & \mathbf{0}_{l\times q_1}\end{bmatrix} =: f_z(\mathbf{A}, \boldsymbol{\varepsilon}_{-Y}).\end{aligned}$$

Assumption 2.1.(d) and Assumption 2.1.(c) together with the assumption that the data matrices consist of row-wise i.i.d. copies of the system variables, yield that $\mathbf{A} \perp\!\!\!\perp \boldsymbol{\varepsilon}_Y$ and $\boldsymbol{\varepsilon}_{-Y} \perp\!\!\!\perp \boldsymbol{\varepsilon}_Y$. This implies that the conditional distribution of $\boldsymbol{\varepsilon}_Y$ given $\mathbf{A}$ and $\boldsymbol{\varepsilon}_{-Y}$ satisfies $P_{\boldsymbol{\varepsilon}_Y|\mathbf{A}=A,\boldsymbol{\varepsilon}_{-Y}=e} = P_{\boldsymbol{\varepsilon}_Y}$ for $P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}$-almost all $(A, e) \in \mathbb{R}^{n\times q} \times \mathbb{R}^{n\times(d+r)}$. Hence, conditional on $\mathbf{A} = A$ and $\boldsymbol{\varepsilon}_{-Y} = e$ we have that $\mathbf{Y}|(\mathbf{A} = A, \boldsymbol{\varepsilon}_{-Y} = e) \overset{a.s.}{=} f_y(A, e) + \boldsymbol{\varepsilon}_Y$, and $\mathbf{Z}|(\mathbf{A} = A, \boldsymbol{\varepsilon}_{-Y} = e) \overset{a.s.}{=} f_z(A, e)$. Now let $(P_\mathbf{A}\mathbf{Z})^+ = (\mathbf{Z}^\intercal P_\mathbf{A}\mathbf{Z})^{-1}\mathbf{Z}^\intercal P_\mathbf{A}$ and $\mathbf{Z}^+ = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$ denote the pseudo-inverse matrices of the almost surely full column rank matrices $P_\mathbf{A}\mathbf{Z}$ and $\mathbf{Z}$. Furthermore, note that the pseudo-inverses are unique for all matrices, i.e., if $P_\mathbf{A}\mathbf{Z} \neq \mathbf{Z}$, then $(P_\mathbf{A}\mathbf{Z})^+ \neq \mathbf{Z}^+$. We realize that $\hat{\alpha}^n_{\text{TSLS}} = (\mathbf{Z}^\intercal P_\mathbf{A}\mathbf{Z})^{-1}\mathbf{Z}^\intercal P_\mathbf{A}\mathbf{Y} = (P_\mathbf{A}\mathbf{Z})^+\mathbf{Y}$ and $\hat{\alpha}^n_{\text{OLS}} = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{Y} = \mathbf{Z}^+\mathbf{Y}$. Thus, with slight abuse of notation we let $Z := f_z(A, e)$ for any A, e, and note that

$$\begin{aligned}&P(\hat{\alpha}^n_{\text{TSLS}} = \hat{\alpha}^n_{\text{OLS}}) \\ &= P((P_\mathbf{A}\mathbf{Z})^+\mathbf{Y} = \mathbf{Z}^+\mathbf{Y}) \\ &= \int P\left([(P_\mathbf{A}\mathbf{Z})^+ - \mathbf{Z}^+]\mathbf{Y} = 0 | \mathbf{A} = A, \boldsymbol{\varepsilon}_{-Y} = e\right) \mathrm{d}P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}(A, e) \\ &= \int P\left([(P_A Z)^+ - Z^+](f_y(A, e) + \boldsymbol{\varepsilon}_Y) = 0\right) \mathrm{d}P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}(A, e) \\ &= \int \mathbb{1}_{(P_A Z \neq Z)} P\left([(P_A Z)^+ - Z^+](f_y(A, e) + \boldsymbol{\varepsilon}_Y) = 0\right) \mathrm{d}P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}(A, e), \qquad \text{(A.20)}\end{aligned}$$

where $P_A = A(A^\intercal A)^{-1}A^\intercal \in \mathbb{R}^{n\times n}$. The last equality is due to the claim that $\mathbb{1}_{(P_A Z \neq Z)} = 1$ for $P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}$ almost all (A, e), or equivalently

$$\int \mathbb{1}_{(P_A Z = Z)} \mathrm{d}P_{\mathbf{A},\boldsymbol{\varepsilon}_{-Y}}(A, e) = \int \mathbb{1}_{(P_\mathbf{A}\mathbf{Z} = \mathbf{Z})} \mathrm{d}P = P(P_\mathbf{A}\mathbf{Z} = \mathbf{Z}) = 0.$$

We prove this claim now.

> We now prove that $P(P_\mathbf{A}\mathbf{Z} = \mathbf{Z}) = 0$. First we note that $P_\mathbf{A}\mathbf{Z} = \mathbf{Z}$ implies that $\mathcal{R}(\mathbf{Z}) \subseteq \mathcal{R}(\mathbf{A})$. Since $\mathbf{Z} = [\mathbf{X}_* \ \mathbf{A}_*]$ with $\mathbf{A} = [\mathbf{A}_* \ \mathbf{A}_{-*}]$ it must hold that $\mathcal{R}(\mathbf{X}_*) \subseteq \mathcal{R}(\mathbf{A})$. Assume without loss of generality that $\mathbf{X}_1$, the first column of $\mathbf{X}$, is also a column of $\mathbf{X}_*$. Note that $\mathcal{R}(\mathbf{X}_*) \subseteq \mathcal{R}(\mathbf{A})$ implies that $\mathbf{X}_1$ can be written as a linear combination of the columns in $\mathbf{A}$, i.e., there exists a $b = (b_1, ..., b_q) \in \mathbb{R}^q$ such that $\mathbf{X}_1 = b_1\mathbf{A}_1 + \cdots + b_q\mathbf{A}_q = \mathbf{A}b$, namely $b = (\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}\mathbf{X}_1$. The marginal reduced form structural equation for $\mathbf{X}_1$ is given by $\mathbf{X}_1 = \mathbf{A}\Pi_{X_1} + \boldsymbol{\varepsilon}\Gamma_{X_1}^{-1} = \mathbf{A}\Pi_{X_1} + \boldsymbol{\varepsilon}_{-Y}\Gamma_{-Y,X_1}^{-1} = \mathbf{A}\Pi_{X_1} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{\varepsilon}} := \boldsymbol{\varepsilon}_{-Y}\Gamma_{-Y,X_1}^{-1}$. These two equalities are only possible if $\tilde{\boldsymbol{\varepsilon}} \in \mathcal{R}(\mathbf{A})$. Note that $\tilde{\boldsymbol{\varepsilon}}$ has jointly

independent marginals (i.i.d. observations). Each coordinate is an independent copy of a linear combination of $1+d+r$ independent random variables $\varepsilon_1, ..., \varepsilon_{1+d+r}$ all with density with respect to Lebesgue measure. We conclude that $\tilde{\boldsymbol{\varepsilon}}$ has density with respect to the n-dimensional Lebesgue measure as the linear combination is non-vanishing. This holds because $\Gamma^{-1}_{-Y,X_1} \neq 0$ by virtue of being a column of the invertible matrix Γ^{-1}, where we have removed the first entry (which was a zero element). Furthermore, since $\mathbf{A} \perp\!\!\!\perp \boldsymbol{\varepsilon}$, we also have that $\mathbf{A} \perp\!\!\!\perp \tilde{\boldsymbol{\varepsilon}}$. Hence, the conditional distribution of $\tilde{\boldsymbol{\varepsilon}}$ given $\mathbf{A}$ satisfies $P_{\tilde{\boldsymbol{\varepsilon}}|\mathbf{A}=A} = P_{\tilde{\boldsymbol{\varepsilon}}}$ for $P_{\mathbf{A}}$-almost all $A \in \mathbb{R}^{n\times q}$. We conclude that

$$\begin{aligned} P(P_{\mathbf{A}}\mathbf{Z} = \mathbf{Z}) &\leq P(\tilde{\boldsymbol{\varepsilon}} \in \mathcal{R}(\mathbf{A})) \\ &= \int P(\tilde{\boldsymbol{\varepsilon}} \in \mathcal{R}(\mathbf{A}) | \mathbf{A} = A) \, \mathrm{d}P_{\mathbf{A}}(A) \\ &= \int P(\tilde{\boldsymbol{\varepsilon}} \in \mathcal{R}(A)) \, \mathrm{d}P_{\mathbf{A}}(A) \\ &= 0. \end{aligned}$$

The last equality follows from the fact that $q = \operatorname{rank}(\mathbf{A}^\intercal\mathbf{A}) = \operatorname{rank}(\mathbf{A}) < n$ implies that $\mathcal{R}(\mathbf{A})$ is a q-dimensional subspace of $\mathbb{R}^n$. Hence, for $P_{\mathbf{A}}$-almost all $A \in \mathbb{R}^{n\times q}$ it holds that $\mathcal{R}(A)$ is a q-dimensional subspace of $\mathbb{R}^n$. The probability that $\tilde{\boldsymbol{\varepsilon}}$ lies in a q-dimensional subspace of $\mathbb{R}^n$ is zero, since it has density with respect to the n-dimensional Lebesgue measure.

Thus, it suffices to show that

$$P\left([(P_A Z)^+ - Z^+](f_y(A, e) + \boldsymbol{\varepsilon}_Y) = 0\right) = 0,$$

for any $A \in \mathbb{R}^{n\times q}$ and $Z = f_z(A, e) \in \mathbb{R}^{n\times(d_1+q_1)}$ with $P_A Z \neq Z$. Therefore, let $A \in \mathbb{R}^{n\times q}$ and $Z = f_z(A, e) \in \mathbb{R}^{n\times(d_1+q_1)}$ with $P_A Z \neq Z$. It holds that $(P_A Z)^+ \neq Z^+$, which implies that $(P_A Z)^+ - Z^+ \neq 0$. Furthermore, we have that

$$[(P_A Z)^+ - Z^+](f_y(A, e) + \boldsymbol{\varepsilon}_Y) = 0,$$

if and only if

$$\boldsymbol{\varepsilon}_{\mathbf{Y}} \in \ker((P_A Z)^+ - Z^+) - [(P_A Z)^+ - Z^+] f_y(A, e),$$

so it suffices to show that $\boldsymbol{\varepsilon}_Y$ has zero probability to be in the affine (translated) subspace

$$\ker((P_A Z)^+ - Z^+) - [(P_A Z)^+ - Z^+] f_y(A, e) \subseteq \mathbb{R}^n.$$

This affine subspace has dimension n if and only if $(P_A Z)^+ - Z^+ = 0$, which we know is false. Hence, the dimension of the affine subspace is strictly less than n. As $\boldsymbol{\varepsilon}_Y$ has density with respect to the n-dimensional Lebesgue measure, we know that the probability of being in a $N < n$ dimensional affine subspace is zero.

Hence, we have shown that $P(\hat{\alpha}^n_{\text{TSLS}} = \hat{\alpha}^n_{\text{OLS}}) = 0$. Combining all of our observations we get that $P(\hat{\alpha}^n_{\text{K}}(\lambda) = \hat{\alpha}^n_{\text{TSLS}}) \leq P(\hat{\alpha}^n_{\text{TSLS}} = \hat{\alpha}^n_{\text{OLS}}) = 0$. We conclude that

$$P\left(\hat{\alpha}^n_{\text{TSLS}} \neq \hat{\alpha}^n_{\text{K}}(\lambda)\right) = 1, \quad \text{for all } \lambda \geq 0.$$

However, we can easily strengthen this to $P(\cap_{\lambda \geq 0}(\hat{\alpha}^n_{\text{TSLS}} \neq \hat{\alpha}^n_{\text{K}}(\lambda))) = 1$. To this end, let ω be a realization in the almost sure set $\cap_{\lambda \in \mathbb{Q}_+} W_\lambda$. Then, $\omega \in \bigcap_{\lambda \geq 0}\left(\hat{\alpha}^n_{\text{TSLS}} \neq \hat{\alpha}^n_{\text{K}}(\lambda)\right)$. Otherwise, there exists an $\tilde{\lambda} \in \mathbb{R}_+ \setminus \mathbb{Q}_+$ such that $\hat{\alpha}^n_{\text{TSLS}} = \hat{\alpha}^n_{\text{K}}(\tilde{\lambda})$. By Lemma 2.6 we have that $\lambda \mapsto l^n_{\text{IV}}(\hat{\alpha}^n_{\text{IV}}(\lambda))$ is monotonically decreasing, but since $\hat{\alpha}^n_{\text{K}}(\tilde{\lambda})$ already minimizes the l^n_{IV} function, so will all $\hat{\alpha}^n_{\text{K}}(\lambda)$ for all $\lambda \geq \tilde{\lambda}$. As $\hat{\alpha}^n_{\text{TSLS}}$ is the unique point that minimizes l^n_{IV} we conclude that $\hat{\alpha}^n_{\text{TSLS}} = \hat{\alpha}^n_{\text{K}}(\lambda)$ for all $\lambda \geq \tilde{\lambda}$, which yields a contradiction. We conclude that $P\left(\cap_{\lambda \geq 0}\left(\hat{\alpha}^n_{\text{TSLS}} \neq \hat{\alpha}^n_{\text{K}}(\lambda)\right)\right) = 1$.

□

Proof of Lemma 2.4: Let Assumption 2.6 and Assumption 2.7 hold, i.e., that $\mathbf{Z}^\intercal\mathbf{Z}$ and $\mathbf{A}^\intercal\mathbf{Z}$ are of full rank and $[\mathbf{Z}\ \mathbf{Y}]$ is of full column rank. Furthermore, let Assumption 2.9 hold, i.e., that $\hat{\alpha}^n_{\text{K}}(\lambda) \notin \mathcal{M}_{\text{IV}}$ for all $\lambda \geq 0$. It holds that (Primal.t.n) has a unique solution and satisfies Slater's condition for all $t > \min_\alpha l^n_{\text{IV}}(\alpha)$ (Lemma 2.2). Furthermore, (Dual.λ.n) has a unique solution for all $\lambda \geq 0$ (Proposition 2.1).

First consider an arbitrary $t \in D_{\text{Pr}}$ and note that the dual problem of (Primal.t.n), not to be confused with the problem (Dual.λ.n), is given by

$$\begin{array}{ll} \text{maximize}_\lambda & g_t(\lambda) \\ \text{subject to} & \lambda \geq 0. \end{array} \tag{A.21}$$

However, (Dual.λ.n) is equivalent with the infimum problem in the definition of $g_t : \mathbb{R}_+ \to \mathbb{R}$ given by

$$g_t(\lambda) := \inf_\alpha \{l^n_{\text{OLS}}(\alpha) + \lambda(l^n_{\text{IV}}(\alpha) - t)\}.$$

Now consider $\hat{\alpha}^n_{\text{Pr}}(t)$ solving the primal (Primal.t.n). Slater's condition is satisfied, so there exists a $\lambda(t) \geq 0$ solving the dual problem and strong duality holds, $l^n_{\text{OLS}}(\hat{\alpha}^n_{\text{Pr}}(t)) = g_t(\lambda(t))$. We will now show that $\hat{\alpha}^n_{\text{Pr}}(t)$ also solves to the K-class penalized regression problem (Dual.$\lambda(t)$.n). That is, we will show that $\hat{\alpha}^n_{\text{Pr}}(t) = \operatorname{argmin}_\alpha\ l^n_{\text{OLS}}(\alpha) + \lambda(t) l^n_{\text{IV}}(\alpha)$. To that end, note that

$$\begin{aligned} g_t(\lambda(t)) &= \inf_\alpha \{l^n_{\text{OLS}}(\alpha) + \lambda(t)(l^n_{\text{IV}}(\alpha) - t)\} = \inf_\alpha \{l^n_{\text{OLS}}(\alpha) + \lambda(t) l^n_{\text{IV}}(\alpha)\} - \lambda(t)t \\ &\leq l^n_{\text{OLS}}(\hat{\alpha}^n_{\text{Pr}}(t)) + \lambda(t)(l^n_{\text{IV}}(\hat{\alpha}^n_{\text{Pr}}(t)) - t) = l^n_{\text{OLS}}(\hat{\alpha}^n_{\text{Pr}}(t)) = g_t(\lambda(t)), \end{aligned}$$

where in the last equality we used strong duality and the second last equality we used that for any constraint bound $t \in D_{\text{Pr}}$ the inequality constraint of (Primal.t.n) is active in the solution $\hat{\alpha}^n_{\text{Pr}}(t)$, i.e., $l^n_{\text{IV}}(\hat{\alpha}^n_{\text{Pr}}(t)) = t$; see Lemma A.4 of Appendix A.6.

Thus, it holds that

$$\begin{aligned} & \inf_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \lambda(t)(l^n_{\mathrm{IV}}(\alpha) - t)\} = l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t)) + \lambda(t)(l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t)) - t) \\ \iff & \inf_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \lambda(t) l^n_{\mathrm{IV}}(\alpha)\} = l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{Pr}}(t)) + \lambda(t) l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t)), \end{aligned}$$

proving that $\hat{\alpha}^n_{\mathrm{Pr}}(t)$ coincides with the unique solution $\hat{\alpha}^n_{\mathrm{K}}(\lambda(t))$ to the K-class problem (Dual.$\lambda(t).n$) as it attains the same objective. Furthermore, there can only be one $\lambda(t)$ solving the dual problem in Equation (A.21). If there are two distinct solutions $\lambda', \lambda'' \geq 0$ with $\lambda' \neq \lambda''$, then by the above observations we get that $\hat{\alpha}^n_{\mathrm{Pr}}(t) = \hat{\alpha}^n_{\mathrm{K}}(\lambda') = \hat{\alpha}^n_{\mathrm{K}}(\lambda'')$, in contradiction to Corollary 2.1.

Conversely, fix $\lambda \geq 0$ and recall that $\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ solves the penalized K-class regression problem (Dual.$\lambda.n$), that is, $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \arg\min_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \lambda l^n_{\mathrm{IV}}(\alpha)\}$. Now consider a primal constraint bound $t(\lambda) := l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ and consider the corresponding primal optimization problem (Primal.$t(\lambda).n$) and its dual form given by

$$\begin{array}{lll} \textit{Primal}: & \text{minimize} & l^n_{\mathrm{OLS}}(\alpha) \\ & \text{subject to} & l^n_{\mathrm{IV}}(\alpha) \leq t(\lambda) \end{array} \qquad \begin{array}{lll} \textit{Dual}: & \text{maximize} & g_{t(\lambda)}(\gamma) \\ & \text{subject to} & \gamma \geq 0, \end{array} \tag{A.22}$$

where $g_{t(\lambda)} : [0, \infty) \to \mathbb{R}$ is given by $g_{t(\lambda)}(\gamma) = \inf_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \gamma[l^n_{\mathrm{IV}}(\alpha) - t(\lambda)]\}$. Here we note that the proposed primal problem satisfies Slater's condition. To see this note that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) \notin \mathcal{M}_{\mathrm{IV}}$, by Assumption 2.9, hence $\inf_{\alpha} l^n_{\mathrm{IV}}(\alpha) = \min_{\alpha} l^n_{\mathrm{IV}}(\alpha) < t(\lambda) = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$. Furthermore, we conclude that $t(\lambda) \in (\min_{\alpha} l^n_{\mathrm{IV}}(\alpha), l_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})] = D_{\mathrm{Pr}}$ as $\lambda \mapsto l_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ is monotonically decreasing and $\hat{\alpha}^n_{\mathrm{K}}(0) = \hat{\alpha}^n_{\mathrm{OLS}}$; see Lemma 2.6.

Let $p^\star$ and $d^\star$ denote the optimal objective values for the above primal and dual problem in Equation (A.22), respectively. It holds that $\hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is primal feasible since it satisfies the inequality constraint of the primal problem in Equation (A.22). This implies that $p^\star \leq l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ since $p^\star$ is the infimum of all attainable objective values. By the non-negative duality gap we also have that

$$\begin{aligned} p^\star &\geq d^\star \\ &= \sup_{\gamma \geq 0} g_{t(\lambda)}(\gamma) \\ &\geq g_{t(\lambda)}(\lambda) \\ &= \inf_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \lambda[l^n_{\mathrm{IV}}(\alpha) - t(\lambda)]\} \\ &= \inf_{\alpha}\{l^n_{\mathrm{OLS}}(\alpha) + \lambda l^n_{\mathrm{IV}}(\alpha)\} - \lambda t(\lambda) \\ &= l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) + \lambda[l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) - l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))] \\ &= l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)), \end{aligned}$$

implying that $l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) = p^\star$. This proves that strong duality holds and that the solution to the K-class regression problem $\alpha^n_{\mathrm{K}}(\lambda)$ solves the primal optimization problem (Primal.$t(\lambda).n$), since it attains the unique optimal objective value while also satisfying the inequality constraint. □

Proof of Theorem 2.3: Fix any $p_{\min} \in (0,1)$ and let Assumptions 2.6 and 2.7 hold, i.e., that $\mathbf{Z}^\intercal\mathbf{Z}$ and $\mathbf{A}^\intercal\mathbf{Z}$ are of full rank and $[\mathbf{Z}\ \mathbf{Y}]$ is of full column rank. Furthermore, let Assumption 2.9 hold, i.e., that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) \notin \mathcal{M}_{\mathrm{IV}}$ for all $\lambda \geq 0$. It holds that (Primal.t.n) has a unique solution and satisfies Slater's condition for all $t > \min_\alpha l^n_{\mathrm{IV}}(\alpha)$ (Lemma 2.2), that (Dual.λ.n) has a unique solution for all $\lambda \geq 0$ (Proposition 2.1) and that $\{\hat{\alpha}^n_{\mathrm{Pr}}(t) : t \in D_{\mathrm{Pr}}\} = \{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\}$ (Lemma 2.4). Finally, we assume that $\lambda^\star_n(p_{\min}) < \infty$. To simplify notation, we write $Q = Q_{\chi^2_q}(1 - p_{\min})$.

We claim that the PULSE estimator can be represented in the dual form of the primal PULSE problem. That is, as a K-class estimator $\hat{\alpha}^n_{\mathrm{PULSE}}(p_{\min}) = \hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))$ with stochastic penalty parameter given by $\lambda^\star_n(p_{\min}) := \inf\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q_{\chi^2_q}(1 - p_{\min})\}$. We show this by proving that $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) = \hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min})))$, which by Theorem 2.2 implies that the claim is true, if the conditions $t^\star_n(p_{\min}) > -\infty$ and $T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min}))) \leq Q$ can be verified from the assumption that $\lambda^\star_n(p_{\min}) < \infty$. First, we note that if $\lambda^\star_n(p_{\min}) < \infty$, then $t^\star_n(p_{\min}) > -\infty$.

> This follows by noting that, with $t(\lambda) := l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$, proof of Lemma 2.4 *ii)* yields that $\hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{Pr}}(t(\lambda))$ for any $\lambda \geq 0$ which yields $\lambda^\star_n(p_{\min}) = \inf\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{Pr}} \circ t(\lambda)) \leq Q\}$. Hence, if $\lambda^\star_n(p_{\min}) < \infty$ we know there exists a $\lambda' \geq 0$ such that $T_n(\hat{\alpha}^n_{\mathrm{Pr}} \circ t(\lambda')) \leq Q$, i.e., there exists a $t' = t(\lambda') \in (\min_{\alpha'} l^n_{\mathrm{IV}}(\alpha'), \infty)$ such that $T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t')) \leq Q$. We have excluded that $t' = \min_{\alpha'} l^n_{\mathrm{IV}}(\alpha')$ as $t' = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda')) > \min_{\alpha'} l^n_{\mathrm{IV}}(\alpha')$ since $\hat{\alpha}_{\mathrm{K}}(\lambda') \notin \mathcal{M}_{\mathrm{IV}}$. Furthermore, we can without loss of generality assume that $t' \in (\min_{\alpha'} l^n_{\mathrm{IV}}(\alpha'), l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})]$ because if $t' > l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})$, then it holds that $T_n(\hat{\alpha}^n_{\mathrm{Pr}}(l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}}))) \leq Q$ as $\hat{\alpha}^n_{\mathrm{Pr}}(l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})) = \hat{\alpha}^n_{\mathrm{Pr}}(t')$ since the ordinary least square solution solves all (Primal.t.n) with constraints bounds larger than $l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})$. As a consequence, the set for which we take the supremum over in the definition of $t^\star_n(p_{\min})$ is non-empty, such that $t^\star_n(p_{\min}) > -\infty$.

Next we show that $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) = \hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min})))$. When this equality is shown, then the remaining condition that $T_n(\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min}))) \leq Q$ follows by Lemma A.2 and we are done. For any constraint bound $t \in D_{\mathrm{Pr}} = (\min_\alpha l^n_{\mathrm{IV}}(\alpha), l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})]$, consider the primal and corresponding dual optimization problems

$$\textit{Primal}: \begin{array}{ll} \text{minimize} & l^n_{\mathrm{OLS}}(\alpha) \\ \text{subject to} & l^n_{\mathrm{IV}}(\alpha) \leq t \end{array} \qquad \textit{Dual}: \begin{array}{ll} \text{maximize} & g_t(\lambda) \\ \text{subject to} & \lambda \geq 0, \end{array} \tag{A.23}$$

with dual function $g_t : \mathbb{R}_{\geq 0} \to \mathbb{R}$ given by $g_t(\lambda) := \inf_\alpha\{l^n_{\mathrm{OLS}}(\alpha) + \lambda(l^n_{\mathrm{IV}}(\alpha) - t)\}$. The proof of Lemma 2.4 yields that there exists a unique $\tilde{\lambda}(t) \geq 0$ solving the dual problem of Equation (A.23) such that

$$\hat{\alpha}^n_{\mathrm{Pr}}(t) = \hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda}(t)).$$

We now prove that $D_{\mathrm{Pr}} \ni t \mapsto \tilde{\lambda}(t)$ is strictly decreasing.

Note that by the definition of g_t and Proposition 2.1 (or equivalently the discussion in the beginning of Section 2.3.5) we have that

$$\begin{aligned} g_t(\lambda) &= \inf_\alpha \{l^n_{\mathrm{OLS}}(\alpha) + \lambda l^n_{\mathrm{IV}}(\alpha)\} - \lambda t \\ &= l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) + \lambda l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) - \lambda t. \end{aligned} \tag{A.24}$$

For any t_1, t_2 with $0 \leq \min_\alpha l^n_{\mathrm{IV}}(\alpha) < t_1 < t_2 \leq l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})$ we have that $g_{t_1}(\tilde{\lambda}(t_1)) \geq g_{t_1}(\tilde{\lambda}(t_2))$ and $g_{t_2}(\tilde{\lambda}(t_2)) \geq g_{t_2}(\tilde{\lambda}(t_1))$ as $\tilde{\lambda}(t)$ maximizes g_t. Hence, by bounding the first term we get that

$$\begin{aligned} g_{t_1}(\tilde{\lambda}(t_1)) - g_{t_2}(\tilde{\lambda}(t_2)) &\geq g_{t_1}(\tilde{\lambda}(t_2)) - g_{t_2}(\tilde{\lambda}(t_2)) \\ &= \tilde{\lambda}(t_2)(t_2 - t_1), \end{aligned} \tag{A.25}$$

where the last equality follows from the representation in Equation (A.24). Similarly, by bounding the other term we get that

$$\begin{aligned} g_{t_1}(\tilde{\lambda}(t_1)) - g_{t_2}(\tilde{\lambda}(t_2)) &\leq g_{t_1}(\tilde{\lambda}(t_1)) - g_{t_2}(\tilde{\lambda}(t_1)) \\ &= \tilde{\lambda}(t_1)(t_2 - t_1). \end{aligned} \tag{A.26}$$

Combining the inequalities from Equations (A.25) and (A.26) we conclude that $\tilde{\lambda}(t_2)(t_2 - t_1) \leq \tilde{\lambda}(t_1)(t_2 - t_1)$ which implies $\tilde{\lambda}(t_2) \leq \tilde{\lambda}(t_1)$, proving that $D_{\mathrm{Pr}} \ni t \mapsto \tilde{\lambda}(t)$, the dual solution as a function of the primal problem constraint bound, is weakly decreasing. We now strengthen this statement to strictly decreasing. For any constraint bound $t \in D_{\mathrm{Pr}} = (\min_\alpha l^n_{\mathrm{IV}}(\alpha), l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})]$ we have that the solution $\hat{\alpha}^n_{\mathrm{Pr}}(t)$ yields an active inequality constraint in the (Primal.$t.n$) problem, i.e., $l^n_{\mathrm{IV}}(\hat{\alpha}_{\mathrm{Pr}}(t)) = t$; see Lemma A.4 of Appendix A.6. Therefore, for any $\min_\alpha l^n_{\mathrm{IV}}(\alpha) < t_1 < t_2 \leq l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{OLS}})$ we get that $l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda}(t_1))) = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_1)) = t_1 < t_2 = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{Pr}}(t_2)) = l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda}(t_2)))$, proving that $\tilde{\lambda}(t_1) \neq \tilde{\lambda}(t_2)$, which implies that $D_{\mathrm{Pr}} \ni t \mapsto \tilde{\lambda}(t)$ is strictly increasing.

Recall, by Lemma 2.4 that the K-class estimators for $\kappa \in [0, 1)$ coincides with the collection of solutions to every primal problem satisfying Slater's condition. That is,

$$\{\hat{\alpha}^n_{\mathrm{K}}(\lambda) : \lambda \geq 0\} = \{\hat{\alpha}^n_{\mathrm{Pr}}(t) : t \in D_{\mathrm{Pr}}\} = \{\hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda}(t)) : t \in D_{\mathrm{Pr}}\}, \tag{A.27}$$

where $\tilde{\lambda}$ is as introduced above.

It now only remains to show that $\tilde{\lambda}(t^\star_n(p_{\min})) = \lambda^\star_n(p_{\min})$, which implies the wanted conclusion as $\hat{\alpha}^n_{\mathrm{Pr}}(t^\star_n(p_{\min})) = \hat{\alpha}^n_{\mathrm{K}}(\tilde{\lambda}(t^\star_n(p_{\min}))) = \hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))$. We know that $\hat{\alpha}^n_{\mathrm{K}} \circ \tilde{\lambda}(t) = \hat{\alpha}^n_{\mathrm{Pr}}(t)$, hence for all $t \in D_{\mathrm{Pr}}$, $(T_n \circ \hat{\alpha}^n_{\mathrm{K}} \circ \tilde{\lambda})(t) = (T_n \circ \hat{\alpha}^n_{\mathrm{Pr}})(t)$, and that for any $A \subseteq [0, \infty)$ it holds that $\tilde{\lambda}(\tilde{\lambda}^{-1}(A)) = A \cap \mathcal{R}(\tilde{\lambda})$, where $\mathcal{R}(\tilde{\lambda}) = \{\tilde{\lambda}(t) : t \in D_{\mathrm{Pr}}\} \subseteq [0, \infty)$ is the range of the reparametrization function $\tilde{\lambda} : D_{\mathrm{Pr}} \to [0, \infty)$. In fact, $\tilde{\lambda}$ is surjective. To see this, note that $[0, \infty) \ni \lambda \mapsto \hat{\alpha}^n_{\mathrm{K}}(\lambda)$ is injective by Corollary 2.1. Thus, $\mathcal{R}(\tilde{\lambda}) = [0, \infty)$ must hold, for otherwise Equation (A.27)

would not hold. Hence, by surjectivity of $\tilde{\lambda}$ we get that for $A \subseteq [0, \infty)$ it holds that $\tilde{\lambda}(\tilde{\lambda}^{-1}(A)) = A$.

Now consider $\hat{\alpha}^n_{\mathrm{Pr}} : D_{\mathrm{Pr}} \to \mathbb{R}^{d_1+q_1}$, $\hat{\alpha}^n_{\mathrm{K}} : [0, \infty) \to \mathbb{R}^{d_1+q_1}$ and $\tilde{\lambda} : D_{\mathrm{Pr}} \to [0, \infty)$ as measurable (which follows by continuity and monotonicity) mappings such that

$$t^\star_n(p_{\min}) = \sup\{(T_n \circ \hat{\alpha}^n_{\mathrm{Pr}})^{-1}(-\infty, Q]\} = \sup\{(T_n \circ \hat{\alpha}^n_{\mathrm{K}} \circ \tilde{\lambda})^{-1}(-\infty, Q]\}.$$

Since $t \mapsto \tilde{\lambda}(t)$ is strictly decreasing, we get that

$$\begin{aligned}
\tilde{\lambda}(t^\star_n(p_{\min})) &= \tilde{\lambda}(\sup\{(T_n \circ \hat{\alpha}^n_{\mathrm{K}} \circ \tilde{\lambda})^{-1}(-\infty, Q]\}) \\
&= \inf\left\{\tilde{\lambda}\left((T_n \circ \hat{\alpha}^n_{\mathrm{K}} \circ \tilde{\lambda})^{-1}(-\infty, Q]\right)\right\} \\
&= \inf\{\tilde{\lambda}(\tilde{\lambda}^{-1}((T_n \circ \hat{\alpha}^n_{\mathrm{K}})^{-1}(-\infty, Q]))\} \\
&= \inf\left\{(T_n \circ \hat{\alpha}^n_{\mathrm{K}})^{-1}(-\infty, Q]\right\} \\
&= \inf\{\lambda \geq 0 : T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq Q\} \\
&= \lambda^\star_n(p_{\min}),
\end{aligned}$$

□

Proof of Lemma 2.5: Let $p_{\min} \in (0, 1)$ and let Assumptions 2.6, 2.7 and 2.9 hold. We have that

$$l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \geq l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{OLS}}) = n^{-1}\|\mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{OLS}}\|_2^2 = n^{-1}\|\mathbf{Y} - P_{\mathbf{Z}}\mathbf{Y}\|_2^2 > 0,$$

as $P_{\mathbf{Z}}\mathbf{Y} \neq \mathbf{Y}$ (by Assumption 2.7 we have that $\mathbf{Y} \notin \mathrm{span}(\mathbf{Z})$, such that the projection of $\mathbf{Y}$ onto the column space of $\mathbf{Z}$ does not coincide with $\mathbf{Y}$ itself). Hence, $T_n : \mathbb{R}^{d_1+q_1} \to \mathbb{R}$ is well-defined, and the following upper bound

$$T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) = n\frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))} \leq n\frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{OLS}})},$$

is valid for every $\lambda \geq 0$. In the under- and just-identified setup we know that there exists an $\tilde{\alpha} \in \mathcal{M}_{\mathrm{IV}} \subseteq \mathbb{R}^{d_1+q_1}$ such that $0 = l^n_{\mathrm{IV}}(\tilde{\alpha})$. Now let $\Lambda > 0$ be given by

$$\Lambda := n\frac{l^n_{\mathrm{OLS}}(\tilde{\alpha})}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{OLS}})Q_{\chi^2_q}(1 - p_{\min})}. \tag{A.28}$$

For any $\lambda > \Lambda$ we have by the non-negativity of $l^n_{\mathrm{OLS}}(\alpha)/\lambda$ that

$$\begin{aligned}
l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) &\leq \lambda^{-1}l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) + l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) = \min_\alpha\{\lambda^{-1}l^n_{\mathrm{OLS}}(\alpha) + l^n_{\mathrm{IV}}(\alpha)\} \\
&\leq \lambda^{-1}l^n_{\mathrm{OLS}}(\tilde{\alpha}) + l^n_{\mathrm{IV}}(\tilde{\alpha}) < \frac{l^n_{\mathrm{OLS}}(\tilde{\alpha})}{\Lambda} = \frac{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{OLS}})Q_{\chi^2_q}(1 - p_{\min})}{n},
\end{aligned}$$

This implies

$$T_n(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \leq n\frac{l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))}{l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{OLS}})} < Q_{\chi^2_q}(1 - p_{\min}),$$

whenever $\lambda > \Lambda$, proving that $\lambda_n^\star(p_{\min}) = \inf\{\lambda \geq 0 : T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda)) \leq Q_{\chi_q^2}(1 - p_{\min})\} < \infty$. Now consider the over-identified setup ($q > d_1 + q_1$). We claim that $\lambda_n^\star(p_{\min}) < \infty$ if and only if $T_n(\hat{\alpha}_{\mathrm{TSLS}}^n) < Q_{\chi_q^2}(1 - p_{\min})$. If $T_n(\hat{\alpha}_{\mathrm{TSLS}}^n) < Q_{\chi_q^2}(1 - p_{\min})$, then by continuity of $\lambda \mapsto \hat{\alpha}_{\mathrm{K}}^n(\lambda)$ and $\alpha \mapsto T_n(\alpha)$ it must hold that $\lambda_n^\star(p_{\min}) < \infty$. This follows by noting that

$$T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda)) \downarrow T_n(\lim_{\lambda\to\infty} \hat{\alpha}_{\mathrm{K}}^n(\lambda)) = T_n(\hat{\alpha}_{\mathrm{TSLS}}^n) < Q_{\chi_q^2}(1 - p_{\min}),$$

when $\lambda \to \infty$, as $\lambda \mapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ is strictly decreasing (Lemma 2.6) Here, we also used that

$$\begin{aligned}
\lim_{\lambda\to\infty} \hat{\alpha}_{\mathrm{K}}^n(\lambda) &= \lim_{\lambda\to\infty} (\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{I} + \lambda P_{\mathbf{A}})\mathbf{Y} \\
&= \lim_{\lambda\to\infty} (\mathbf{Z}^\intercal(\lambda^{-1}\mathbf{I} + P_{\mathbf{A}})\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\lambda^{-1}\mathbf{I} + P_{\mathbf{A}})\mathbf{Y} \\
&= (\mathbf{Z}^\intercal P_{\mathbf{A}}\mathbf{Z})^{-1}\mathbf{Z}^\intercal P_{\mathbf{A}}\mathbf{Y} \\
&= \hat{\alpha}_{\mathrm{TSLS}}^n.
\end{aligned}$$

Hence, there must exist a $\lambda \in [0, \infty)$ such that $T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda)) < Q_{\chi_q^2}(1 - p_{\min})$, proving that $\lambda_n^\star(p_{\min}) < \infty$. Furthermore, note that the above arguments also imply that $T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda)) > T_n(\hat{\alpha}_{\mathrm{TSLS}}^n)$, for any $\lambda \geq 0$, as $\lambda \mapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ is strictly decreasing and $T_n(\hat{\alpha}_{\mathrm{TSLS}}^n)$ is the limit as $\lambda \to \infty$.

Conversely, assume that $\lambda_n^\star(p_{\min}) < \infty$, which implies that there exists a $\lambda' \in [0, \infty)$ such that $T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda')) \leq Q_{\chi_q^2}(1 - p_{\min})$. Thus,

$$T_n(\hat{\alpha}_{\mathrm{TSLS}}^n) < T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda')) \leq Q_{\chi_q^2}(1 - p_{\min}),$$

proving that the converse implication also holds.

We furthermore note that, if the acceptance region is empty, that is

$$\mathcal{A}_n(1 - p_{\min}) := \{\alpha \in \mathbb{R}^{d_1+q_1} : T_n(\alpha) \leq Q_{\chi_q^2}(1 - p_{\min})\} = \emptyset,$$

then it obviously holds that $\lambda_n^\star(p_{\min}) = \{\lambda \geq 0 : T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda)) \leq Q_{\chi_q^2}(1 - p_{\min})\} = \infty$. The possibility of the acceptance region being empty, follows from the fact that the Anderson-Rubin confidence region can be empty; see Remark A.2. To realize that the Anderson-Rubin confidence region can be empty we refer to the discussions and Monte-Carlo simulations of Davidson and MacKinnon (2014).

□

Proof of Lemma 2.7: Assume that $\lambda_n^\star(p_{\min}) < \infty$ and that Assumption 2.6.(a) and Assumption 2.7 hold. Consider Algorithm A.1 for any fixed $N \in \mathbb{N}$. The first 'while loop' guarantees that $\lambda_{\min}$ and $\lambda_{\max}$ are such that $\lambda^\star \in (\lambda_{\min}, \lambda_{\max}]$. This is seen by noting that $\lambda \mapsto T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda))$ is monotonically decreasing (Lemma 2.6) and that $\lambda_n^\star(p_{\min}) < \infty$. Hence, $T_n(\hat{\alpha}_{\mathrm{K}}^n(\lambda_{\max}))$ eventually drops below $Q_{\chi_q^2}(1 - p)$. The second 'while loop' keeps iterating until the interval $(\lambda_{\min}, \lambda_{\max}]$, which is guaranteed to contain $\lambda_n^\star(p_{\min})$, has a length less than or equal to $1/N$. Let $\lambda_{\min}$

and $\lambda_{\max}$ denote the last boundaries achieved before the procedure terminates. Then $0 \leq \text{Binary.Search}(N,p) - \lambda_n^\star(p_{\min}) = \lambda_{\max} - \lambda_n^\star(p_{\min}) \leq \lambda_{\max} - \lambda_{\min} \leq 1/N$. Hence, $\text{Binary.Search}(N,p) - \lambda_n^\star(p_{\min}) \to 0$, as $N \to \infty$. □

Proof of Theorem 2.4: Consider the just- or over-identified setup ($q \geq d_1 + q_1$), let Assumption 2.4 hold. We furthermore assume that the population rank condition, Assumption 2.8, i.e., $E(AZ^\intercal)$ is of full rank, are satisfied. We furthermore work under the finite-sample conditions of Assumptions 2.6 and 2.7, i.e., that $\mathbf{A}^\intercal\mathbf{A}$ and $\mathbf{Z}^\intercal\mathbf{A}$ are of full rank and $\mathbf{Y} \notin \text{span}(\mathbf{Z})$ for all sample-sizes $n \in \mathbb{N}$ almost surely. The first two of these are not strictly necessary as the population version of these rank assumptions guarantee that $\mathbf{A}^\intercal\mathbf{A}$ and $\mathbf{Z}^\intercal\mathbf{A}$ are of full rank with probability tending to one; see proof of Proposition 2.2. Likewise, we can drop the last finite-sample assumption as it is almost surely guaranteed if we assume that the distribution of ε_Y has density with respect to Lebesgue measure. The proof below is easily modified to accommodate these more relaxed assumptions, but for notational simplicity we prove the statement under the stronger finite-sample assumptions. We also let Assumption 2.9 hold which in addition with the previous assumptions guarantees that the dual representation of the PULSE holds whenever $\lambda_n^\star(p_{\min}) < \infty$; see Theorem 2.3. Furthermore, many of the previous theorems and lemmas were shown for a specific realization that satisfies the finite sample assumptions. Hence, we may only invoke the conclusions of these theorems almost surely. Note that the assumptions guarantee that the TSLS estimator is consistent, i.e., $\hat{\alpha}^n_{\text{TSLS}} \xrightarrow{P} \alpha_0$.

Fix any $p_{\min} \in (0,1)$ and let an arbitrary $\varepsilon > 0$ be given. We want to prove that $P(\|\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) - \alpha_0\| > \varepsilon) \to 0$. To that end, define the events $(A_n)_{n\in\mathbb{N}}$ by $A_n := (T_n(\hat{\alpha}^n_{\text{TSLS}}) < Q_{\chi^2_q}(1-p_{\min}))$, such that

$$\begin{aligned} P(\|\hat{\alpha}^n_{\text{PULSE+}}(p_{\min}) - \alpha_0\| > \varepsilon) =& P((\|\hat{\alpha}^n_{\text{PULSE}}(p_{\min}) - \alpha_0\| > \varepsilon) \cap A_n) && \text{(A.29)} \\ &+ P((\|\hat{\alpha}^n_{\text{ALT}} - \alpha_0\| > \varepsilon) \cap A_n^c), && \text{(A.30)} \end{aligned}$$

for all $n \in \mathbb{N}$. The last term, Equation (A.30), tends to zero as $n \to \infty$,

$$P((\|\hat{\alpha}^n_{\text{ALT}} - \alpha_0\| > \varepsilon) \cap A_n^c) \leq P(\|\hat{\alpha}^n_{\text{ALT}} - \alpha_0\| > \varepsilon) \to 0,$$

by the assumption that $\hat{\alpha}^n_{\text{ALT}} \xrightarrow{P} \alpha_0$ as $n \to \infty$. In regards to the first term, the right-hand side of Equation (A.29), we note that $A_n = (\lambda_n^\star(p_{\min}) < \infty)$, by Lemma 2.5. Formally, this event equality only holds when intersecting both sides with the almost sure event that the finite sample rank condition holds. However, we suppress this intersection for ease of notation. Thus, on A_n, it holds that $\hat{\alpha}^n_{\text{PULSE}}(p_{\min}) = \hat{\alpha}^n_{\text{K}}(\lambda_n^\star(p_{\min}))$, by Theorem 2.3, implying that

$$P((\|\hat{\alpha}^n_{\text{PULSE}}(p_{\min}) - \alpha_0\| > \varepsilon) \cap A_n) = P((\|\hat{\alpha}^n_{\text{K}}(\lambda_n^\star(p_{\min})) - \alpha_0\| > \varepsilon) \cap A_n).$$

Furthermore, Lemma A.2 yields that on A_n, it holds that

$$T_n(\hat{\alpha}^n_{\text{K}}(\lambda_n^\star(p_{\min}))) \leq Q_{\chi^2_q}(1-p_{\min}),$$

or equivalently

$$l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) \leq n^{-1} Q_{\chi^2_q}(1-p_{\min}) l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})).$$

On A_n, the stochastic factor in the upper bound above, is further bounded from above by

$$\begin{aligned} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min}))) &\leq \sup_{\lambda \geq 0} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \\ &= \lim_{\lambda \to \infty} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda)) \\ &= l^n_{\mathrm{OLS}}(\lim_{\lambda \to \infty} \hat{\alpha}^n_{\mathrm{K}}(\lambda)) \\ &= l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{TSLS}}), \end{aligned}$$

where we used continuity of $\alpha \mapsto l^n_{\mathrm{OLS}}(\alpha)$, that $\lambda \mapsto l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda))$ is weakly increasing (Lemma 2.6) and that $\lim_{\lambda\to\infty} \hat{\alpha}^n_{\mathrm{K}}(\lambda) = \hat{\alpha}^n_{\mathrm{TSLS}}$. Recall that the TSLS estimator is consistent $\hat{\alpha}^n_{\mathrm{TSLS}} \xrightarrow{P} \alpha_0$, where α_0 is the causal coefficient of Z onto Y. Hence, Slutsky's theorem and the weak law of large numbers yield that

$$\begin{aligned} l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{TSLS}}) &= n^{-1}(\mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}})^\intercal(\mathbf{Y} - \mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}}) \\ &= n^{-1}\mathbf{Y}^\intercal\mathbf{Y} + (\hat{\alpha}^n_{\mathrm{TSLS}})^\intercal n^{-1}\mathbf{Z}^\intercal\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}} - 2n^{-1}\mathbf{Y}^\intercal\mathbf{Z}\hat{\alpha}^n_{\mathrm{TSLS}} \\ &\xrightarrow{P} E(Y^2) + \alpha_0^\intercal E(ZZ^\intercal)\alpha_0 - 2E(YZ^\intercal)\alpha_0 \\ &= E[(Y - Z\alpha_0)^2]. \end{aligned}$$

Thus, on the event A_n, we have that

$$0 \leq l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) \leq n^{-1} Q_{\chi^2_q}(1-p_{\min}) l^n_{\mathrm{OLS}}(\hat{\alpha}^n_{\mathrm{TSLS}}) =: H_n,$$

where the upper bound H_n converges to zero in probability by Slutsky's theorem. Furthermore, note that

$$\begin{aligned} l^n_{\mathrm{IV}}(\alpha_0) &= \|n^{-1/2}(\mathbf{A}^\intercal\mathbf{A})^{-1/2}\mathbf{A}^\intercal(\mathbf{Y} - \mathbf{Z}\alpha_0)\|_2^2 \\ &= \|(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1/2} n^{-1}\mathbf{A}^\intercal\mathbf{U}_Y\|_2^2 \\ &\xrightarrow{P} \|E(AA^\intercal)^{-1/2}E(AU_Y)\|_2^2 \\ &= \|E(AA^\intercal)^{-1/2}E(A)E(U_Y)\|_2^2 \\ &= 0, \end{aligned}$$

where we used that $Y = Z^\intercal\alpha_0 + U_Y$, Assumption 2.4.(a): $A \perp\!\!\!\perp U_Y$, and Assumption 2.4.(b): $E(A) = 0$ (Alternatively, $E(U_Y|A) = 0$).

Now define a sequence of (everywhere) well-defined estimators $(\tilde{\alpha}_n)_{n\in\mathbb{N}}$ by

$$\tilde{\alpha}_n := \mathbb{1}_{A_n}\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) + \mathbb{1}_{A_n^c}\alpha_0,$$

for each $n \in \mathbb{N}$. We claim that the loss function l^n_{IV} evaluated in this estimator tends to zero in probability, i.e., as $n \to \infty$ it holds that

$$l^n_{\mathrm{IV}}(\tilde{\alpha}_n) = \|(n^{-1}\mathbf{A}^\intercal\mathbf{A})^{-1/2} n^{-1}\mathbf{A}^\intercal(\mathbf{Y} - \mathbf{Z}\tilde{\alpha}_n)\|_2^2 \xrightarrow{P} 0. \tag{A.31}$$

This holds by the above observations as for any $\varepsilon' > 0$ we have that

$$\begin{aligned} P(|l^n_{\mathrm{IV}}(\tilde{\alpha}_n)| > \varepsilon') &= P((|l^n_{\mathrm{IV}}(\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})))| > \varepsilon') \cap A_n) \\ &\quad + P((|l^n_{\mathrm{IV}}(\alpha_0)| > \varepsilon') \cap A^c_n) \\ &\leq P((|H_n| > \varepsilon') \cap A_n) + P((|l^n_{\mathrm{IV}}(\alpha_0)| > \varepsilon') \cap A^c_n) \\ &\leq P(|H_n| > \varepsilon') + P(|l^n_{\mathrm{IV}}(\alpha_0)| > \varepsilon') \to 0, \end{aligned}$$

when $n \to \infty$. Now define the random linear maps $g_n : \Omega \times \mathbb{R}^{d_1+q_1} \to \mathbb{R}^q$ by

$$g_n(\omega, \alpha) := (n^{-1}\mathbf{A}^\intercal(\omega)\mathbf{A}(\omega))^{-1/2} n^{-1}\mathbf{A}^\intercal(\omega)\mathbf{Z}(\omega)\alpha,$$

for all $n \in \mathbb{N}$. The maps (g_n) converge point-wise, that is, for each α, in probability to $g : \mathbb{R}^{d_1+q_1} \to \mathbb{R}^q$, given by $g(\alpha) := E(AA^\intercal)^{-1/2}E(AZ^\intercal)\alpha$, as $n \to \infty$. The map g is injective. This follows by Assumption 2.1.(h) and Assumption 2.8, which implies and state that $E(AA^\intercal) \in \mathbb{R}^{q\times q}$ and $E(AZ^\intercal) \in \mathbb{R}^{q\times(d_1+q_1)}$ are of full rank, respectively, hence $\operatorname{rank}(E(AA^\intercal)^{-1/2}E(AZ^\intercal)) = \operatorname{rank}(E(AZ^\intercal)) = d_1 + q_1$, since we are in the just- and over-identified setup, where $q \geq d_1 + q_1$. We conclude that g is injective, as its matrix representation is of full column rank. Furthermore, by Equation (A.31) it holds that $g_n(\tilde{\alpha}_n) \xrightarrow{P} E(AA^\intercal)^{-1/2}E(AY)$. Hence, we have that

$$\begin{aligned} g_n(\tilde{\alpha}_n) - g_n(\alpha_0) &\xrightarrow{P} E(AA^\intercal)^{-1/2}E(AY) - E(AA^\intercal)^{-1/2}E(AZ^\intercal)\alpha_0 \\ &= E(AA^\intercal)^{-1/2}E(AU_Y) \\ &= 0, \end{aligned}$$

as $n \to \infty$. Lemma A.3 of Appendix A.6 now yields that $\tilde{\alpha}_n \xrightarrow{P} \alpha_0$. Finally, note that as $\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) = \tilde{\alpha}_n$ on A_n we have that

$$\begin{aligned} P((\|\hat{\alpha}^n_{\mathrm{K}}(\lambda^\star_n(p_{\min})) - \alpha_0\| > \varepsilon) \cap A_n) &= P((\|\tilde{\alpha}_n - \alpha_0\| > \varepsilon) \cap A_n) \\ &\leq P(\|\tilde{\alpha}_n - \alpha_0\| > \varepsilon) \\ &\to 0, \end{aligned}$$

proving that $\hat{\alpha}^n_{\mathrm{PULSE+}}(p_{\min}) \xrightarrow{P} \alpha_0$, as $n \to \infty$. □

A.6. Auxiliary Lemmas

Lemma A.3. *Suppose that $g_n : \mathbb{R}^G \to \mathbb{R}^K$ are random linear maps converging point-wise in probability to a non-random linear map $g : \mathbb{R}^G \to \mathbb{R}^K$ that is injective. If*

$$g_n(\hat{\beta}_n - \beta_0) \xrightarrow[n\to\infty]{P} 0,$$

then $\hat{\beta}_n$ is a consistent estimator of β_0. That is, $\hat{\beta}_n \xrightarrow[n\to\infty]{P} \beta_0$.

Proof of Lemma A.3: As g is injective, we have that $\text{rank}(g) = G$, and as such $\text{rank}(g^\intercal g) = (g) = G$ which implies that $g^\intercal g$ is invertible. Furthermore, by Slutsky's theorem we get that $g_n \xrightarrow{P} g \implies g_n^\intercal g_n \xrightarrow{P} g^\intercal g$, as $n \to \infty$, that is, for any $\varepsilon > 0$,

$$P(||g_n^\intercal g_n - g^\intercal g|| \leq \varepsilon) = P(g_n^\intercal g_n \in \overline{B(g^\intercal g, \varepsilon)}) \underset{n\to\infty}{\to} 1.$$

Here $\|\cdot\|$ is any norm on the set of $G \times G$ matrices and $\overline{B(g^\intercal g, \varepsilon)}$ is the closed ball around $g^\intercal g$ with radius ε with respect the this norm. Now note that the set NS_G of all non-singular $G \times G$ matrices is an open subset of all $G \times G$ matrices, which implies that there exists an $\varepsilon > 0$, such that $\overline{B(g^\intercal g, \varepsilon)} \subseteq \text{NS}_G$. Hence, $g_n^\intercal g_n$ is invertible with probability tending towards 1, that is, $P(g_n^\intercal g_n \in \text{NS}_G) \underset{n\to\infty}{\to} 1$. Let $h_n : \Omega \to \text{NS}_G$ be given by

$$h_n(\omega) := \mathbb{1}_{(g_n^\intercal g_n \in \text{NS}_G)} g_n^\intercal(\omega) g_n(\omega) + \mathbb{1}_{(g_n^\intercal g_n \in \text{NS}_G)^c} I.$$

Then $h_n \xrightarrow[n\to\infty]{P} g^\intercal g$, since for any $\varepsilon > 0$

$$\begin{aligned} P(\|h_n - g^\intercal g\| > \varepsilon) &= P((\|g_n^\intercal g_n - g^\intercal g\| > \varepsilon) \cap (g_n^\intercal g_n \in \text{NS}_G)) \\ &\quad + P((\|I - g^\intercal g\| > \varepsilon) \cap (g_n^\intercal g_n \in \text{NS}_G)^c) \\ &\leq P(\|g_n^\intercal g_n - g^\intercal g\| > \varepsilon) + P(g_n^\intercal g_n \in \text{NS}_G)^c) \\ &\underset{n\to\infty}{\to} 0, \end{aligned}$$

Continuity of the inverse operator and the continuous mapping theorem, yield that $\|h_n^{-1}\|_{\text{op}} \xrightarrow{P} \|(g^\intercal g)^{-1}\|_{\text{op}} \in \mathbb{R}$ and $\|g_n^\intercal\|_{\text{op}} \xrightarrow{P} \|g^\intercal\|_{\text{op}} \in \mathbb{R}$ as n tends to infinity, where $\|\cdot\|_{\text{op}}$ is the operator norm induced by the Euclidean norm $\|\cdot\|_2$. Furthermore,

$$\begin{aligned} \|g_n^\intercal g_n(\hat{\beta}_n - \beta_0)\|_2 &\leq \|g_n^\intercal\|_{\text{op}} \|g_n(\hat{\beta}_n - \beta_0)\|_2 \\ &\xrightarrow[n\to\infty]{P} \|g^\intercal\|_{\text{op}} \cdot 0 \\ &= 0, \end{aligned}$$

by the assumptions and Slutsky's theorem. Hence, for any $\varepsilon > 0$

$$\begin{aligned} P(\|h_n(\hat{\beta}_n - \beta_0)\|_2 > \varepsilon) &= P((\|g_n^\intercal g_n(\hat{\beta}_n - \beta_0)\|_2 > \varepsilon) \cap (g_n^\intercal g_n \in \text{NS}_G)) \\ &\quad + P((\|\hat{\beta}_n - \beta_0\|_2 > \varepsilon) \cap (g_n^\intercal g_n \in \text{NS}_G)^c) \\ &\leq P((\|g_n^\intercal g_n(\hat{\beta}_n - \beta_0)\|_2 > \varepsilon)) + P((g_n^\intercal g_n \in \text{NS}_G)^c) \\ &\underset{n\to\infty}{\to} 0. \end{aligned}$$

Thus,

$$\begin{aligned} \|\hat{\beta}_n - \beta_0\|_2 &= \|h_n^{-1} h_n(\hat{\beta}_n - \beta_0)\|_2 \\ &\leq \|h_n^{-1}\|_{\text{op}} \|h_n(\hat{\beta}_n - \beta_0)\|_2 \\ &\xrightarrow[n\to\infty]{P} \|(g^\intercal g)^{-1}\|_{\text{op}} \cdot 0 \\ &= 0, \end{aligned}$$

by Slutsky's theorem, yielding that $\hat{\beta}_n$ is an consistent estimator of β_0. □

Lemma A.4. *Let $\hat{\alpha}$ be a solution to a constrained optimization problem of the form*

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^k}{\text{minimize}} \quad & f(\alpha) \\ \text{subject to} \quad & g(\alpha) \leq c, \end{aligned}$$

where f is an everywhere differentiable strictly convex function on $\mathbb{R}^k$ for which a stationary point exists, g is continuous and $c \in \mathbb{R}$. If the stationary point of f is not feasible, then the constraint inequality is tight (active) in the solution $\hat{\alpha}$, that is, $g(\hat{\alpha}) = c$.

Proof of Lemma A.4: Since $\hat{\alpha}$ feasible and the stationary point of f is not feasible, we know that $\hat{\alpha}$ is not a stationary point of f, hence $Df(\hat{\alpha}) \neq 0$. Now assume that the constraint bound is not tight (active) in the solution $\hat{\alpha}$, that is $g(\hat{\alpha}) < c$. By continuity of g, we know that there exists an $\varepsilon > 0$, such that for all $\alpha \in B(\hat{\alpha}, \varepsilon)$, it holds that $g(\alpha) < c$. Furthermore, since $Df(\hat{\alpha}) = 0$, we can look at the line segment going through $\hat{\alpha}$ in the direction of the negative gradient of f in $\hat{\alpha}$. That is, $l : \mathbb{R} \to \mathbb{R}^k$ defined by $l(t) = \hat{\alpha} - tDf(\hat{\alpha})$. Note that

$$D(f \circ l)(0) = Df(l(0))Dl(0) = -Df(\hat{\alpha})Df(\hat{\alpha})^\intercal = -\|Df(\hat{\alpha})\| < 0,$$

meaning that the derivative of $f \circ l : \mathbb{R} \to \mathbb{R}$ is negative in zero. Therefore, there exists a $\delta > 0$, such that for all $t \in (0, \delta)$ it holds that $f \circ l(t) < f \circ l(0)$, i.e.,

$$f(\hat{\alpha} - tDf(\hat{\alpha})) < f(\hat{\alpha}).$$

Thus, for sufficiently small t', is it holds that $t' < \delta$ and $\hat{\alpha} - t'Df(\hat{\alpha}) \in B(\hat{\alpha}, \varepsilon)$. We conclude that $\tilde{\alpha} := \hat{\alpha} - tDf(\hat{\alpha})$ is feasible, $g(\tilde{\alpha}) < c$, and super-optimal compared to $\hat{\alpha}$, $f(\tilde{\alpha}) < f(\hat{\alpha})$, which contradicts that $\hat{\alpha}$ solves the optimization problem. In words, if the solution is not tight we can take a small step in the negative gradient direction of the objective function and get a better objective value while still being feasible. □

A.7. Additional Remarks

Remark A.1 (Model misspecification)**.** Theorem 2.1 still holds under the following three model misspecifications, which may arise from erroneous non-sample information and unobserved endogenous variables (these violations may break the identification of $\alpha_{0,*}$ and generally render the K-class estimators inconsistent even when P-lim $\kappa = 1$).

(a) Exclude included endogenous variables. Consider the setup where no hidden variable enters the target equation given by $Y = \gamma_0^\intercal X + \beta_0^\intercal A + \varepsilon_Y$, with $\varepsilon_Y \perp\!\!\!\perp A$. If we erroneously exclude an endogenous variable that directly affects Y, i.e., $\gamma_{0,-*} \neq 0$, this is equivalent to drawing inference from the model $Y = \gamma_{0,*}^\intercal X_* + \beta_{0,*}^\intercal A_* + U$, where $U = \varepsilon_Y + \gamma_{0,-*}^\intercal X_{-*}$. If $E(A_{-*}U) = E(A_{-*}X_{-*}^\intercal)\gamma_{0,-*} \neq 0$, we have introduced dependence that renders at least some of the instruments A_{-*} invalid, breaking identifiability.

(b) Exclude included exogenous variables. Consider again the setup from (a) where there is no hidden variables entering the target equation. If we erroneously exclude a exogenous variable that directly affects Y, i.e., $\beta_{0,-*} \neq 0$, then this is equivalent with drawing inference from the model $Y = \gamma_{0,*}^\intercal X_* + \beta_{0,*}^\intercal A_* + U$, where $U = \varepsilon_Y + \beta_{0,-*}^\intercal A_{-*}$. It holds that $E(A_{-*}U) = E(A_{-*}A_{-*}^\intercal)\beta_{0,-*} \neq 0$, again rendering the instruments invalid.

(c) Possibility of hidden endogenous variables. Consider the case with included hidden variables that are directly influenced by the excluded exogenous variables, i.e., $A_{-*} \to H \to Y$. This implies that the excluded exogenous variables A_{-*} are correlated with the collapsed noise variable in the structural equation $Y = \alpha_{0,*}^\intercal Z_* + U$, where $U = \varepsilon_Y + \eta_0^\intercal H$ with $\eta_0 \neq 0$. In the case that $E(A_{-*}U) = E(A_{-*}H^\intercal)\eta_0 \neq 0$ the instruments are invalid.

∘

Remark A.2 (Connection to the Anderson-Rubin Test)**.** Our acceptance region $\mathcal{A}_n^c(1-p_{\min}) := \{\alpha \in \mathbb{R}^{d_1+q_1} : T_n^c(\alpha) \leq Q_{\chi_q^2}(1-p_{\min})\}$, is closely related to the Anderson-Rubin (Anderson and Rubin, 1949) confidence region of the simultaneous causal parameter $\alpha_0 = (\gamma_0, \alpha_0)$ in an identified model. When the causal parameter α_0 is identifiable, i.e., in a just- or over-identified setup ($q \geq d_1 + q_2$) and Assumption 2.8 holds, only the causal parameter yields regression residuals $Y - \alpha_0^\intercal Z$ that are uncorrelated with the exogenous variables A. In this restricted setup, our null hypothesis is equivalent with $\tilde{H}_0(\alpha) : \alpha = \alpha_0$. The hypothesis $\tilde{H}_0(\alpha)$ can be tested by the Anderson-Rubin test and all non-rejected coefficients constitute the Anderson-Rubin confidence region of α_0, which is given by $\mathrm{CR}_{\mathrm{AR}}^{ex,n}(1-p_{\min}) := \left\{\alpha \in \mathbb{R}^{d_1+q_1} : T_n^{\mathrm{AR}}(\alpha) \leq Q_{F(q,n-q)}(1-p_{\min})\right\}$, where $Q_{F(q,n-q)}(1-p_{\min})$ is the $1-p_{\min}$ quantile of the F distribution with q and $n-q$ degrees of freedom and the Anderson-Rubin test-statistic $T_n^{\mathrm{AR}}(\alpha)$ is given by

$$T_n^{\mathrm{AR}}(\alpha) := \frac{n-q}{q} \frac{(\mathbf{Y} - \mathbf{Z}\alpha)^\intercal P_{\mathbf{A}} (\mathbf{Y} - \mathbf{Z}\alpha)}{(\mathbf{Y} - \mathbf{Z}\alpha)^\intercal P_{\mathbf{A}}^\perp (\mathbf{Y} - \mathbf{Z}\alpha)} = \frac{n-q}{q} \frac{l_{\mathrm{IV}}^n(\alpha)}{l_{\mathrm{OLS}}^n(\alpha) - l_{\mathrm{IV}}^n(\alpha)}.$$

The confidence region $\mathrm{CR}_{\mathrm{AR}}^{ex,n}$ is exact whenever several regularity conditions are satisfied, such as deterministic exogenous variables and normal distributed errors (Anderson and Rubin, 1949, Theorem 3). In a general SEM model the regularity conditions are not fulfilled, but changing the rejection threshold to $Q_{\chi_q^2/q}(1-p_{\min})$,

we obtain an asymptotically valid confidence region. That is,

$$\mathrm{CR}_{\mathrm{AR}}^{as,n}(1-p_{\min}) := \left\{\alpha \in \mathbb{R}^{d_1+q_1} : T_n^{\mathrm{AR}}(\alpha) \leq Q_{\chi^2_q/q}(1-p_{\min})\right\},$$

is an asymptotically valid approximate confidence region (Anderson and Rubin, 1950, Theorem 6). This relies on the fact that $T_n^{\mathrm{AR}}(\alpha) \xrightarrow{\mathcal{D}} \chi^2_q/q$ under the null and T_n^{AR} diverges to infinity under the general alternative. The test-statistic $T_n^c(\alpha)$ can be seen as a scaled coefficient of determination (R^2-statistic) for which $T_n^{\mathrm{AR}}(\alpha)$ is the corresponding F-statistic. That is, one can realize that

$$T_n^{\mathrm{AR}}(\alpha) = \frac{n-q}{q}\frac{T_n^c(\alpha)/c(n)}{1-T_n^c(\alpha)/c(n)} \leq Q_{\chi^2_q/q}(1-p_{\min}),$$

is equivalent to

$$\frac{n-q+Q_{\chi^2_q}(1-p_{\min})}{c(n)} T_n^c(\alpha) \leq Q_{\chi^2_q}(1-p_{\min}).$$

Thus, if $Q_{\chi^2_q}(1-p_{\min}) \geq q$, then $\mathcal{A}_n(1-p_{\min}) \supseteq \mathrm{CR}_{\mathrm{AR}}^{as,n}(1-p_{\min})$ and $\mathcal{A}_n(1-p_{\min}) \subseteq \mathrm{CR}_{\mathrm{AR}}^{as,n}(1-p_{\min})$ otherwise, where $\mathcal{A}_n(1-p_{\min})$ is the acceptance region when using the scaling scheme $c(n) = n$. Furthermore, $\mathcal{A}_n^c(1-p_{\min})$, the acceptance region under a general scaling scheme $c(n) \sim n$, is asymptotically equivalent to the Anderson-Rubin approximate confidence region $\mathrm{CR}_{\mathrm{AR}}^{as,n}(1-p_{\min})$. If we choose the specific scaling to be $c(n) = n - q + Q_{\chi^2_q}(1-p_{\min})$, then they coincide, $\mathrm{CR}_{\mathrm{AR}}^{as,n}(1-p_{\min}) = \mathcal{A}_n^c(1-p_{\min})$ for each $n \in \mathbb{N}$. Whenever the Anderson-Rubin confidence region is exact, we could change the rejection threshold from $Q_{\chi^2_q}(1-p_{\min})$ to $c(n)Q_{B(q/2,(n-q)/2)}(1-p_{\min})$ and also get an exact acceptance region, where $B(q/2,(n-q)/2)$ is the Beta distribution with shape and scale parameter $q/2$ and $(n-q)/2$ respectively. ○

Remark A.3 (Connections to pre-test estimators)**.** It has been suggested to use pre-test for choosing between the TSLS and OLS estimator. When using the Hausman test for endogeneity (Hausman, 1978) one considers the pre-test estimator studied by, e.g., Chmelarova and Hill (2010) and Guggenberger (2010). If H denotes the Hausman test-statistic that rejects the hypothesis of endogeneity when $H \leq Q$, the pre-test estimator is given by $\alpha^n_{\mathrm{pretest}} = 1_{(H\leq Q)}\alpha^n_{\mathrm{OLS}} + 1_{(H>Q)}\alpha^n_{\mathrm{TSLS}}$. The PULSE estimator can be seen as a pre-test estimator using the Anderson-Rubin test as a test for endogeneity. However, PULSE differs from the above in the sense that when endogeneity is not rejected we do not revert to the TSLS estimate but rather to the coefficient within the Anderson-Rubin confidence region that minimizes the mean squared prediction error. ○

A.8. Simulation Study

A.8.1. Distributional Robustness

We first illustrate the distributional robustness property of K-class estimators discussed in Section 2.2.3.3 in a finite sample setting. We consider the model given

by

$$X := A + U_X, \qquad Y := \gamma X + U_Y,$$

where $\gamma = 1$ and $A \sim N(0,1)$ independent of $\binom{U_X}{U_Y} \sim \mathcal{N}((\begin{smallmatrix}0\\0\end{smallmatrix}), (\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}))$. We estimate γ from $n = 2000$ observations generated by the above system and estimate $\hat{\gamma}^n_{\mathrm{K}}(\kappa)$ for all $\kappa \in \{0, 3/4, 1\}$ for which the corresponding population coefficients are given by $\gamma_{\mathrm{K}}(0) = \gamma_{\mathrm{OLS}} = 1.25$, $\gamma_{\mathrm{K}}(3/4) = 1.1$ and $\gamma_{\mathrm{K}}(1) = \gamma_{\mathrm{TSLS}} = 1$. We repeat the simulation 50 times and save the estimated coefficients. Figure A.2 illustrates the distributional robustness property of Theorem 2.1. For all estimated coefficients $\hat{\gamma}$ of γ we have plotted the analytically computed worst case mean squared prediction error (MSPE) under all hard interventions of absolute strength up to x given by

$$\sup_{|v| \leq x} E^{\mathrm{do}(A:=v)}[(Y - \hat{\gamma}X)^2] = x^2(1-\hat{\gamma})^2 + \hat{\gamma}^2 + 3(1-\hat{\gamma}) \tag{A.32}$$

against the maximum intervention strength x for the range $x \in [0, 6]$. The plot also shows results for the population coefficient as seen in Rothenhäusler et al. (2021, Figure 2).

In all 50 repetitions the K-class estimator for $\kappa = 3/4$ outperforms both OLS and TSLS in terms of worst case MSPE for maximum intervention strength of 2. This is in line with the theory presented in Section 2.2.3.3. In terms of population coefficients our theoretical results predict that $\kappa = 3/4$ is worst case MSPE superior, relative to OLS and TSLS, for all maximum intervention strengths in the range $[1.37, 3]$. Among the 50 repetitions we find the outcomes for which the superiority range of $\kappa = 3/4$ has the shortest and longest superiority range length. The shortest superiority range is $[1.27, 2.15]$ and the longest is $[1.46, 5.54]$. Clearly, these numbers vary with changing sample size and number of repetitions. For example, with 50, 200, 500, 2000, 5000 and 10000 observations and 50 repetitions, the median lengths of the MSPE superiority range for $\kappa = 3/4$ equal 0.82, 1.16, 1.44, 1.74, 1.58 and 1.63, respectively (1.63 is also the length of the theoretically computed interval $[1.37, 3]$).

A.8.2. Estimating causal effects

In this subsection we investigate the finite sample behaviour of the PULSE estimator by simulation experiments. We look at how the PULSE estimator fairs in comparison to other well-known single equation estimators in terms of different performance measures. We generate $n \in \mathbb{N}$ realizations of the SEM in question and construct the estimators of interest based on these n observations. This is repeated $N \in \mathbb{N}$ times, allowing us to estimate different finite sample performance measures of the estimators of interest. The characterization of weak instruments through the minimum eigenvalue of G_n, a multivariate analogue to the first stage F-statistic, as introduced in Stock and Yogo (2002) is important for some of our experimental findings. We refer the reader to Appendix A.10 for a brief introduction.

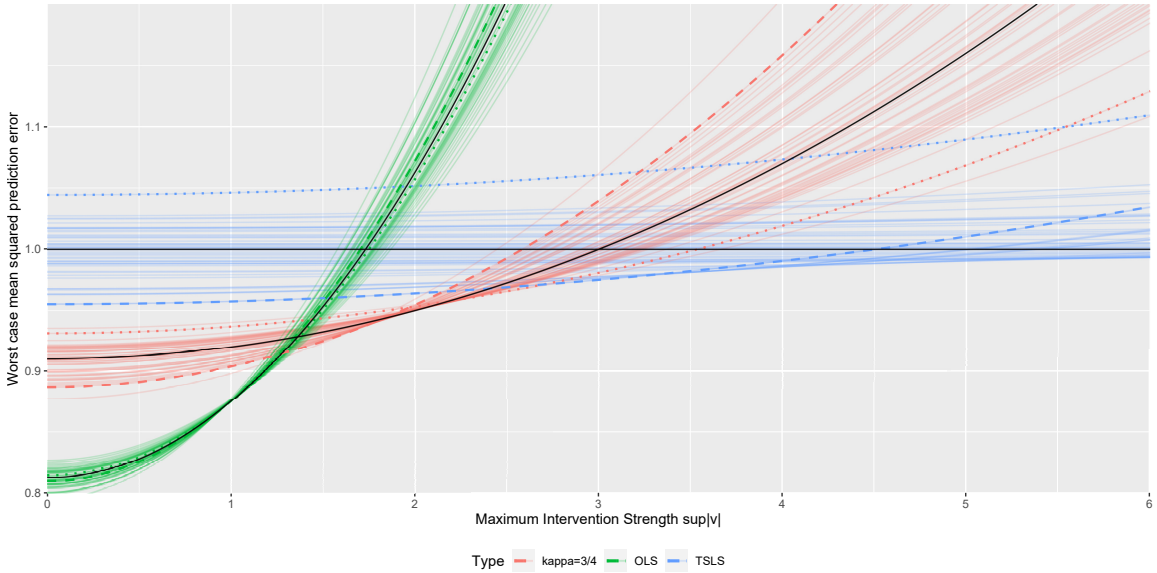


Figure A.2: Distributional Robustness of K-class estimators. The plot shows the worst case MSPE against the maximum intervention strength condsidered. Each of the 50 repetitions corresponds to three lines (green, red, blue), corresponding to the three estimates using $\kappa \in \{0, 3/4, 1\}$, respectively. The solid black line corresponds to the population coefficients. The OLS is optimal for small interventions but yields a large loss for strong interventions; the TSLS is optimal for large interventions but yields a relatively large loss for small interventions. Choosing a κ different from zero and allows us to trade off these two regimes. The dashed and the dotted lines correspond to the two samples, for which the interval on which the $\kappa = 3/4$ estimator outperforms TSLS and OLS in terms of worst case MSPE is shortest and longest, respectively.

A.8.2.1. Benchmark Estimators and Performance Measures

We compare the PULSE(5) estimator, that is PULSE with $p_{\min} = 0.05$, to four specific K-class estimators that are well-known to have second moments (in sufficiently over-identified setups). This will allow us to conduct both bias and mean squared error analysis of estimators. Most importantly, we benchmark against Fuller estimators. The κ-parameter of the Fuller estimators are given by $\kappa^n_{\text{FUL}}(a) = \kappa^n_{\text{LIML}} - \frac{a}{n-q}$, where $n - q$ is the degrees of freedom in the first stage regression, $a > 0$ is a hyper parameter and κ^n_{LIML} is the stochastic κ-parameter corresponding the to LIML estimator. One way to represent the κ-parameter of the LIML estimator is $\kappa_{\text{LIML}} = \lambda_{\min}(W_1 W^{-1})$ where $\lambda_{\min}$ denotes the smallest eigenvalue, W_1 and W are defined as $W = [\mathbf{Y}\ \mathbf{X}]^\intercal P^\perp_{\mathbf{A}} [\mathbf{Y}\ \mathbf{X}]$ and $W_1 = [\mathbf{Y}\ \mathbf{X}]^\intercal P^\perp_{\mathbf{A}_*} [\mathbf{Y}\ \mathbf{X}]$, and $P^\perp_{\mathbf{A}} = \mathbf{I} - \mathbf{A}(\mathbf{A}^\intercal \mathbf{A})^{-1}\mathbf{A}^\intercal$; see, e.g., Amemiya (1985). We choose to benchmark the PULSE estimator against the following K-class estimators: OLS ($\kappa = 0$), TSLS ($\kappa = 1$), Fuller(1) ($\kappa = \kappa^n_{\text{FUL}}(1)$) and Fuller(4)

($\kappa = \kappa^n_{\text{FUL}}(4)$).

The Fuller(1) estimator is approximately unbiased in that the mean bias is zero up to $\mathcal{O}(n^{-2})$ (Fuller, 1977, Theorem 1) and Fuller(4) exhibits approximate superiority in terms of MSE compared to all other Fuller estimators (Fuller, 1977, Corollary 2). As we shall see below the PULSE estimator has good MSE performance when instruments are weak and therefore we especially benchmark against Fuller(4) which has shown better MSE performance than TSLS in simulation studies when instruments are weak; see e.g. Hahn et al. (2004). In the over-identified setup we let the PULSE estimator revert to Fuller(4) whenever the dual representation is infeasible.

We compare the estimators in terms of bias and mean squared error (MSE), which for an n-sample estimator $\hat{\alpha}_n$ with target $\alpha \in \mathbb{R}^{d_1+q_1}$ are given by $\text{Bias}(\hat{\alpha}_n) = E(\hat{\alpha}_n) - \alpha \in \mathbb{R}^{d_1+q_1}$, $\text{MSE}(\hat{\alpha}_n) = E[(\hat{\alpha}_n - \alpha)(\hat{\alpha}_n - \alpha)^\intercal] \in \mathbb{R}^{(d_1+q_1)\times(d_1+q_1)}$. The empirical quantities, estimated from N independent repetitions are denoted by $\widehat{\text{Bias}}(\hat{\alpha}_n)$ and $\widehat{\text{MSE}}(\hat{\alpha}_n)$. In the multivariate setting, we compare biases by comparing their Euclidean norms. When comparing MSEs, we call $\hat{\alpha}_n$ MSE superior to $\tilde{\alpha}_n$ if they are ordered in the partial ordering generated by the proper cone of positive semi-definite matrices (that is, $\widehat{\text{MSE}}(\tilde{\alpha}_n) - \widehat{\text{MSE}}(\hat{\alpha}_n)$ is positive semi-definite). We also consider the ordering of its scalarizations given by the determinant and trace (the latter satisfies $\text{trace}(\widehat{\text{MSE}}(\hat{\alpha}_n)) = \text{trace}(\widehat{\text{Var}}(\hat{\alpha}_n)) + \|\widehat{\text{Bias}}(\hat{\alpha}_n)\|_2^2$).

We conduct the simulation experiments even though it is not proved that the PULSE estimator has finite second moments. In the simulations, the empirical estimates of the mean squared error were stable, possibly even more so than for the Fuller estimators for which we know that second moments exists in settings where the noise is Gaussian; see e.g., Chao et al. (2012); Fuller (1977).

Below we describe two multivariate simulation experiments and refer the reader to Section 2.4.2.1 in the main paper for a univariate simulation experiment.

A.8.2.2. Varying Confounding Multivariate Experiment.

In this simulation scheme we consider just-identified two-dimensional instrumental variable models with the SEM and causal graph illustrated in Figure A.3. Since we want to compare MSE statistics that require estimators with second moments we drop comparisons with the TSLS estimator.

Here, $\xi, \delta \in \mathbb{R}^{2\times 2}$, $\mu \in \mathbb{R}^2$ and (N_A, N_X, N_X, N_Y) are independent noise innovations. We let $\gamma = (0, 0)$ and let the noise innovations for A, H, Y have distribution $(N_A, N_H, N_Y) \sim \mathcal{N}(0, I)$.

We randomly generate 10000 models by letting $N_X \sim \mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{smallmatrix}\right)\right)$, where the standard deviations is drawn by $\sigma_1^2, \sigma_2^2 \sim \text{Unif}(0.1, 1)$ and all other model coefficients are drawn according to

$$\xi_{11}, \xi_{12}, \xi_{21}, \xi_{22}, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}, \mu_1, \mu_2 \sim \text{Unif}(-2, 2).$$

The hidden confounding induces dependence between the collapsed noise variables

$$
\begin{aligned}
A &:= N_A \in \mathbb{R}^2, \\
H &:= N_H \in \mathbb{R}^2, \\
X &:= \xi^\intercal A + \delta^\intercal H + N_X \in \mathbb{R}^2, \\
Y &:= \gamma^\intercal X + \mu^\intercal H + N_Y \in \mathbb{R}.
\end{aligned}
$$

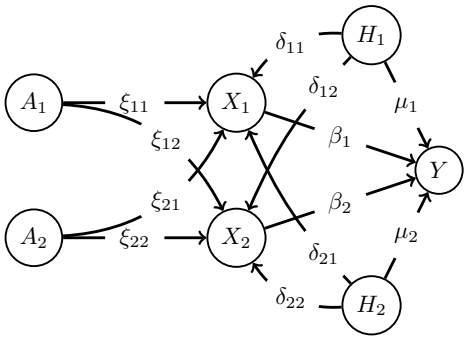


Figure A.3: The SEM and graph representation used for simulating data in the experiments described in Section A.8.2.2.

$U_X = \delta^\intercal H + N_X$ and $U_Y = \mu^\intercal H + N_Y$, which we capture by a normalized cross covariance vector $\rho := \Sigma_{U_X}^{-1/2} \Sigma_{U_X U_Y} \Sigma_{U_Y}^{-1/2} \in \mathbb{R}^2$, where $\Sigma_{U_X} = \mathrm{Var}(U_X)$, $\Sigma_{U_X U_Y} = \mathrm{Cov}(U_X, U_Y)$ and $\Sigma_{U_Y} = \mathrm{Var}(U_Y)$. As such, the degree of confounding can be explained by the norm of ρ given by $\|\rho\|_2^2 = \Sigma_{U_Y U_X} \Sigma_{U_X}^{-1} \Sigma_{U_X U_Y} / \Sigma_{U_Y} = \mu^\intercal \delta (\delta^\intercal \delta + \mathrm{diag}(\sigma_1^2, \sigma_2^2))^{-1} \delta^\intercal \mu / (\mu^\intercal \mu + 1)$. For each of the 10000 generated models we simulate $n = 50$ observations and compute the PULSE and benchmark estimators and repeat this $N = 5000$ times to estimate the performance measures.

Figure A.4 shows the relative change in the determinant and trace of the MSE matrix and the Euclidean norm of the bias vector. Similarly to the univariate setup, PULSE seems to perform better than Fuller(1) and Fuller(4) in terms of the determinant and trace for settings with weak confounding (small $\|\rho\|_2$) and weak instruments (small $\lambda_{\min}(\hat{E}_N G_n)$). Most of the MSE matrices do not allow for an ordering: PULSE is MSE superior to Fuller(1), Fuller(4), and OLS in 9.2%, 4.6% and 1% of the cases, while the MSE matrices are not ordered in 90.8%, 95.4% and 95.8% of the cases. Note that both Fuller(1) and Fuller(4) is never MSE superior to PULSE. In contrast to the univariate setup, there are models with very weak instruments for which Fuller outperforms PULSE; these models seems to be exclusively with strong confounding. We also see models with strong confounding and moderate to strong instrument strength where the PULSE is superior and models with weak confounding where PULSE is inferior. Hence, the degree of confounding $\|\rho\|_2$ does not completely characterize whether or not PULSE is superior to the Fuller estimators in terms of MSE performance measures in the multi-dimensional setting. In regards to the bias we see that both Fuller estimators are less biased than PULSE for all but a few models with very weak instruments. Furthermore, PULSE is for models with strong confounding less biased than OLS but has comparable bias for models with small to moderate confounding.

We also conducted the above simulation experiment for $\gamma = (1, 1)$ and $\gamma = (-1, 1)$. The results (not shown but available in the folder 'Plots' in the code repository) are similar to the case $\gamma = (0, 0)$ and the above observations still apply.

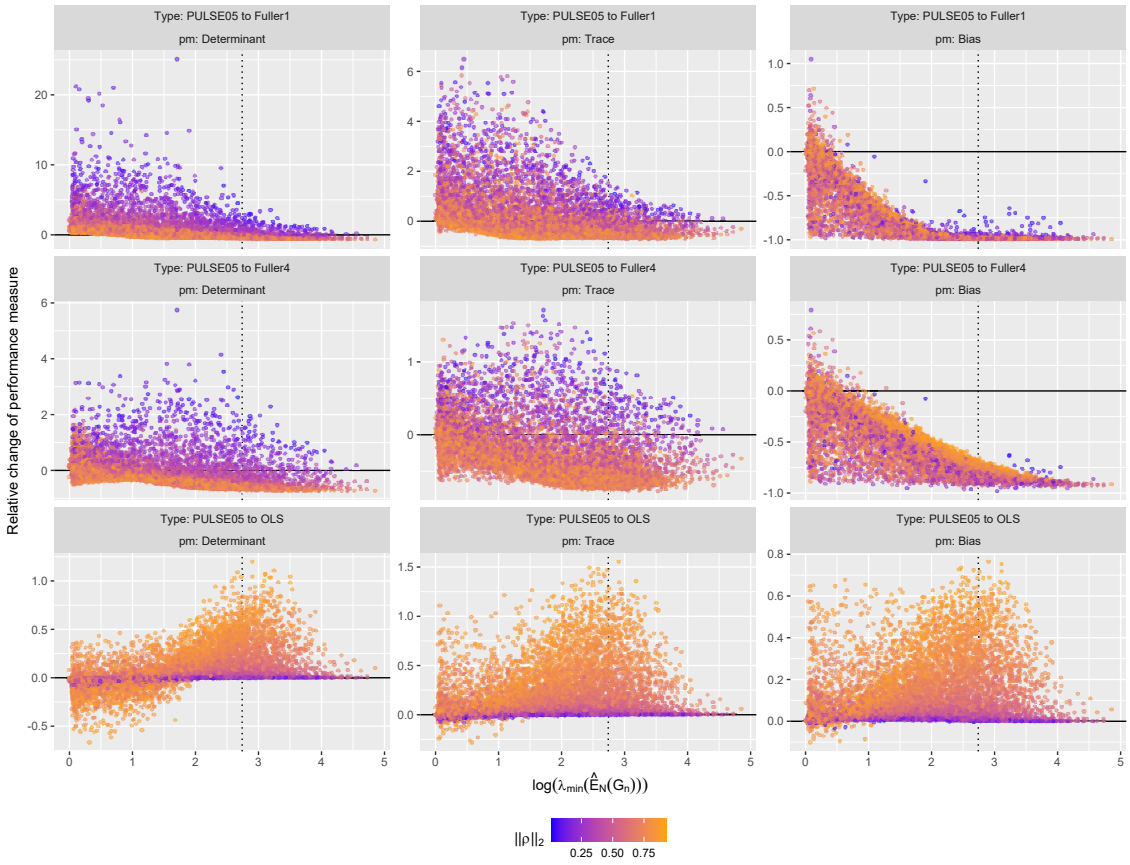


Figure A.4: Illustrations of the relative change in the determinant (left) and trace (middle) of the MSE matrix and the Euclidean norm of the bias vector (right) (a positive relative change means that PULSE is better). Each of the 10000 models corresponds to a point which is color-graded according the the value of $\|\rho\|_2$ (which indicates the strength of confounding), see Section A.8.2.2. PULSE tends to outperform the Fuller estimators for weak instruments and weak confounding. The vertical dotted line at log(15.5) corresponds to a rejection threshold for weak instruments based on relative change in bias for Fuller estimators (Stock and Yogo, 2002, Table 5.3). Note that the lowest possible negative relative change is -1.

Appendix A.11 shows the results of additional experiments, where we consider, e.g., PULSE with $p_{\min} = 0.1$.

A.8.2.3. Fixed Confounding Multivariate Experiment.

In the varying confounding experiment, we saw that when $\|\rho\|_2$ is small then the majority of the simulated models had PULSE superior to Fuller(1) and Fuller(4) in terms of the determinant and trace of MSE. However, we also saw models with large $\|\rho\|_2$ where PULSE was still superior and models with small $\|\rho\|_2$ where PULSE

was inferior. In this experiment, we will investigate this further by fixing the confounding strength $\|\rho\|_2$ and investigating other model aspects that affect which estimator is superior. That is, we consider models with structural assignments given by

$$A := N_A \in \mathbb{R}^2, \quad X := \xi^\intercal A + U_X \in \mathbb{R}^2, \quad Y := \gamma^\intercal X + U_Y \in \mathbb{R},$$

for some $\xi \in \mathbb{R}^{2\times 2}$ and independent noise innovations $(N_A, (U_X, U_Y))$. We let $\gamma = (0, 0)$ and fix the noise innovations for A with distribution $N_A \sim \mathcal{N}(0, I)$. We let

$$\begin{pmatrix} U_X \\ U_Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \eta & \varphi_1 \\ \eta & 1 & \varphi_2 \\ \varphi_1 & \varphi_2 & 1 \end{pmatrix} \right),$$

for some $\eta, \varphi_1, \varphi_2 \in [0, 1)$. With this noise structure we have that $\|\rho\|_2^2 = (\varphi_1^2 + \varphi_2^2 - 2\eta\varphi_1\varphi_2)/(1 - \eta^2)$, and when $\varphi = \varphi_1 = \varphi_2$ it holds that $\|\rho\|_2^2 = 2\varphi^2/(1 + \eta)$. We randomly generate 5000 copies of ξ with each entry drawn by Unif$(-2, 2)$ distribution. For each model, that is, each combination of selected noise-parameter values and ξ, we simulate $n = 50$ observations and compute the estimators. This is repeated $N = 5000$ times to estimate the performance measures.

In Figure A.5 we have illustrated the relative change in the performance measures when comparing PULSE to Fuller(4). For setups with weak confounding ($\|\rho\|_2 = 0.2$), it is seen that if instruments are sufficiently weak ($\lambda_{\min}(\hat{E}_N(G_n)) \leq 15.5$), then PULSE is superior to Fuller(4) in terms of both the determinant and trace performance measures. For setups with larger confounding there are still models where PULSE is superior but the characterization of superiority by weakness of instruments is no longer valid.

In Table A.1 the percentage of models for which PULSE is superior to Fuller(4) in terms of the MSE partial ordering, determinant and trace performance measures is presented. It is seen that setups with identical $\|\rho\|_2$ does not yield similar comparisons between PULSE and Fuller(4).

For any two setups with identical confounding strength $\|\rho\|_2$ we see that decreasing η yields a larger percentage of models for which PULSE is superior in terms of the determinant and trace. Furthermore, we see that decreasing $\|\rho\|_2$ (for fixed η) has a similar effect. Thus, it seems that both ρ and η negatively influences the size of the parameter space of ξ for which PULSE is superior to Fuller(4) in terms of both the determinant and trace performance measures. However, superiority with respect to the partial ordering of the MSE matrices does not exhibit similar behaviour. Decreasing $\|\rho\|_2$ (for fixed η) still leads to a percentage increase but decreasing η (for fixed $\|\rho\|_2$) leads to a percentage decrease, of models for which PULSE is superior to Fuller(4).

Table A.1: MSE superiority

Model Parameters				PULSE Superiority (%)		
$\|\rho\|_2$	η	φ_1	φ_2	MSE	determinant	trace
0.20	0.80	0.19	0.19	48.46	85.52	86.74
0.20	0.20	0.15	0.15	32.34	98.66	92.16
0.50	0.80	0.47	0.47	1.60	13.04	19.80
0.50	0.20	0.39	0.39	0.76	19.86	27.86
0.80	0.80	0.76	0.76	0.14	7.48	12.80
0.80	0.20	0.62	0.62	0.06	7.64	15.50

Note: The rows show different noise-parameter values for the different experimental setups. The last three columns describe the percentage of models (out of the 5000 randomly generated models) for which PULSE (with $p_{\min} = 0.05$) is superior to Fuller(4) in terms of the MSE partial ordering, determinant and trace performance measures. Whenever PULSE is not superior to Fuller(4) in terms of the MSE partial ordering the MSE matrices are not comparable.

A.8.3. Under-identified setup

In an under-identified setup the causal parameter is not identified by instrumental variable methods. Instead the usual two stage least square procedure, $\arg\min_\alpha l_{\mathrm{IV}}(\alpha)$, yields an entire linear solution space of coefficients that renders the regression residuals uncorrelated with the instruments. The causal coefficient lies within this solution space but we are unable to identify it. In the under-identified setup, the population PULSE coefficient is the point in the solution space which provides the best mean squared prediction error. That is, the population PULSE coefficient is given by

$$\alpha^* = \operatorname*{arg\,min}_{\alpha: E[A(Y-Z\alpha)]=0} E[(Y - Z\alpha)^2] = \operatorname*{arg\,min}_{\alpha: l_{\mathrm{IV}}(\alpha)=0} l_{\mathrm{OLS}}(\alpha).$$

The PULSE estimator in the under-identified setup remains unchanged from the exposition in the main paper. Here, the function l^n_{IV} does not have a unique solution but we can define a modified TSLS estimator

$$\hat{\alpha}^n_{\mathrm{TSLS.mod}} := \lim_{\kappa \uparrow 1} \alpha^n_{\mathrm{K}}(\kappa) = \operatorname*{arg\,min}_{\alpha: l^n_{\mathrm{IV}}(\alpha)=0} l^n_{\mathrm{OLS}}(\alpha).$$

The modified TSLS estimator is the minimum of a quadratic function subject to a feasible linear constraint, and can be computed efficiently using QP solvers.

A.8.3.1. Under-indentified Example

Consider an under-identified setup with structural assignments given by

$$\begin{aligned} A &:= \varepsilon_A, \qquad H := \varepsilon_H, \qquad X_1 := \eta A + \delta_1 H + \varepsilon_1, \\ Y &:= \beta X_1 + \delta_2 H + \varepsilon_Y, \qquad X_2 := \gamma Y + \varepsilon_2, \end{aligned}$$

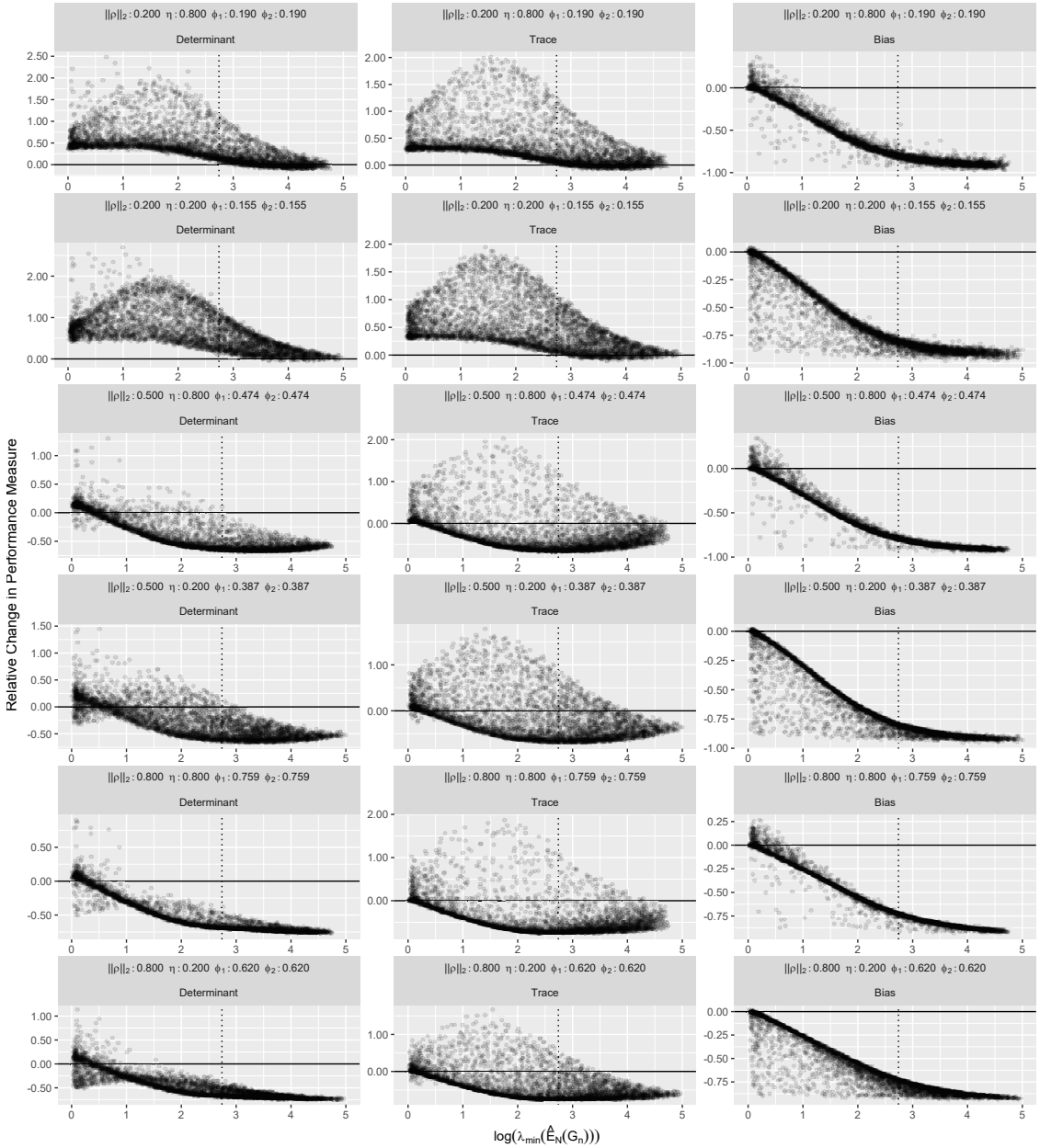


Figure A.5: Illustrations of the relative change from PULSE to Fuller(4) in the determinant and trace of the MSE matrix and the Euclidean norm of the bias vector. The vertical dotted line at log(15.5) corresponds to a rejection threshold for weak instruments based on relative change in bias for Fuller estimators (Stock and Yogo, 2002, Table 5.3).

with $(\varepsilon_A, \varepsilon_H, \varepsilon_Y, \varepsilon_1, \varepsilon_2) \sim \mathcal{N}(0, I_5)$. The causal graph of this structural equation model is illustrated in Figure A.6. In general, the causal parameter β is not identifiable. Existing methods (e.g., Peters et al., 2016; Pfister et al., 2021; Rojas-Carulla et al., 2018b) propose to look for invariant sets that yield residuals which are uncorrelated with A after regressing Y on that set. In general, because of the

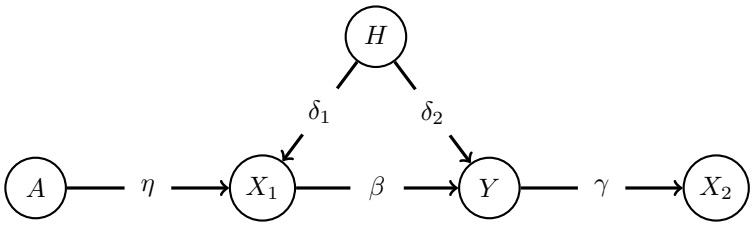


Figure A.6: Causal graph of the under-indentified setup in Section A.8.3.1 Here, H is hidden and the causal parameter β is, in general, not identifiable from the distribution over (A, X_1, X_2, Y). Existing methods in machine learning try to find invariant sets of covariates (i.e., sets S that, after regressing Y on X_S, yield residuals which are uncorrelated with A). In this example, no such set exists. PULSE finds a solution and outputs a vector with non-zero coefficients for X_1 and X_2.

hidden variable H, no such sets exist either. The best predictive model under all invariant models, however, is still well-defined. To see this, let us derive the population PULSE coefficient

$$\alpha^* = \underset{\alpha: l_{\mathrm{IV}}(\alpha)=0}{\arg\min}\, E[(Y - \alpha_1 X_1 - \alpha_2 X_2)^2].$$

We know that a necessary and sufficient condition for $l_{\mathrm{IV}}(\alpha) = 0$ is that $\mathrm{Corr}(Y - \alpha_1 X_1 - \alpha_2 X_2, A) = 0$. We have

$$\begin{aligned} Y - \alpha_1 X_1 - \alpha_2 X_2 &= Y - \alpha_1 X_1 - \alpha_2(\gamma Y + \varepsilon_2) \\ &= (1 - \alpha_2\gamma)(\beta X_1 + \delta_2 H + \varepsilon_Y) - \alpha_1 X_1 - \alpha_2 \varepsilon_2 \\ &= (\beta - \alpha_1 - \alpha_2\gamma\beta) X_1 + (1 - \alpha_2\gamma)\delta_2 H \\ &\quad + (1 - \alpha_2\gamma)\varepsilon_Y - \alpha_2\varepsilon_2. \end{aligned}$$

As $\eta \neq 0$, the regression residuals are uncorreleted with A if and only if $\alpha_1 = (1 - \alpha_2\gamma)\beta$. Hence,

$$\begin{aligned} \alpha^* &= \underset{\alpha: \alpha_1 = (1-\alpha_2\gamma)\beta}{\arg\min}\, E[((1 - \alpha_2\gamma)\delta_2 H + (1 - \alpha_2\gamma)\varepsilon_Y - \alpha_2\varepsilon_2)^2] \\ &= \underset{\alpha: \alpha_1 = (1-\alpha_2\gamma)\beta}{\arg\min}\, (1 - \alpha_2\gamma)^2\delta_2^2 \mathrm{Var}(H) + (1 - \alpha_2\gamma)^2 \mathrm{Var}(\varepsilon_Y) + \alpha_2^2 \mathrm{Var}(\varepsilon_2). \end{aligned}$$

The latter function is convex in α_2, so the minimum is attained in a stationary point. We have that

$$\begin{aligned} &\frac{\partial}{\partial \alpha_2}(1 - \alpha_2\gamma)^2\delta_2^2 \mathrm{Var}(H) + (1 - \alpha_2\gamma)^2 \mathrm{Var}(\varepsilon_Y) + \alpha_2^2 \mathrm{Var}(\varepsilon_2) \\ &= 2\left[\alpha_2(\mathrm{Var}(\varepsilon_2) + \gamma^2\delta_2^2 \mathrm{Var}(H) + \gamma^2 \mathrm{Var}(\varepsilon_Y)) - \gamma\delta_2^2 \mathrm{Var}(H) - \gamma \mathrm{Var}(\varepsilon_Y)\right] \\ &= 0, \end{aligned}$$

if and only if

$$\alpha_2(\mathrm{Var}(\varepsilon_2) + \gamma^2\delta_2^2\mathrm{Var}(H) + \gamma^2\mathrm{Var}(\varepsilon_Y) = \gamma\delta_2^2\mathrm{Var}(H) + \gamma\mathrm{Var}(\varepsilon_Y).$$

Hence,

$$\alpha_2^* = \frac{(\mathrm{Var}(\varepsilon_Y) + \delta_2^2\mathrm{Var}(H))\gamma}{\mathrm{Var}(\varepsilon_2) + (\mathrm{Var}(\varepsilon_Y) + \delta_2^2\mathrm{Var}(H))\gamma^2} = \frac{(1+\delta_2^2)\gamma}{1 + (1+\delta_2^2)\gamma^2}; \tag{A.33}$$

$$\alpha_1^* = (1 - \alpha_2^*\gamma)\beta. \tag{A.34}$$

We now generate models by randomly drawing the model coefficients using $\alpha \sim \mathrm{Unif}(1,2)$, $\delta_1 \sim \mathrm{Unif}(1,2), \delta_2 \sim \mathrm{Unif}(1,2), \gamma \sim \mathrm{Unif}(1,2)$ and $\eta \sim \mathrm{Unif}(0.1,1)$ and compute the corresponding population quantities according to Equation (A.33).

For different sample sizes, we then simulate data sets from such models and compute the PULSE estimator. Figure A.7 shows the trace of the estimated MSE of the PULSE estimator (with $p_{\min} = 0.05$) when comparing to the population quantity derived above. For each model and sample size, the MSE is estimated based on 100 repetitions. As sample size increases, the MSE indeed approaches the population quantity.

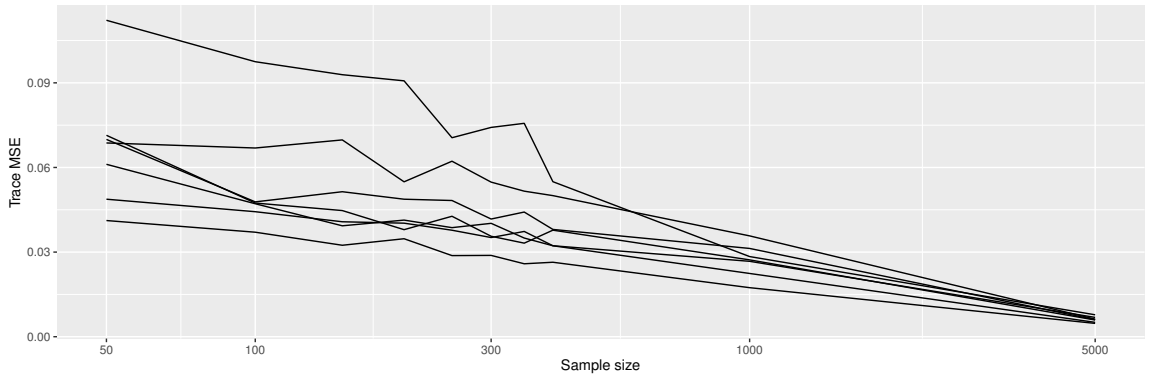


Figure A.7: Illustration of the trace of the estimated MSE matrix of the PULSE estimator in the under-identified setup based on 100 repetitions. PULSE converges towards the population quantities computed in Equation (A.33).

As a comparison, we also implemented the TSLS modification from Equation (A.33). Similarly to the identified setups, the TSLS modification may come with poor finite sample properties, in particular for weak instruments and small sample size. Indeed, in this example we observe that PULSE has superior MSE properties for small sample sizes. For example, the trace MSE for the PULSE estimator is on average (over 1000 random models) 50% lower than the trace MSE of the modified TSLS estimator for a sample size of 50.

A.9. Empirical Applications

We now consider three classical instrumental variable applications (see Albouy (2012) and Buckles and Hungerman (2013) for discussions on the underlying assumptions).

A.9.1 "Does compulsory school attendance affect schooling and earnings?" by Angrist and Krueger (1991). This paper investigates the effects of education on wages. The endogenous effect of education on wages are remedied by instrumenting education on quarter of birth indicators.

A.9.2 "Using geographic variation in college proximity to estimate the return to schooling" by Card (1993). This paper also investigates the effects of education on wages. In this paper education is instrumented by proximity to college indicator.

A.9.3 "The colonial origins of comparative development: An empirical investigation" by Acemoglu et al. (2001). This paper investigates the effects of extractive institutions (proxied by protection against expropriation) on the gross domestic product (GDP) per capita. The endogeneity of the explanatory variables are remedied by instrumenting protection against expropriation on early European settler mortality rates.

For each study, we replicate the OLS and TSLS estimates of these studies and provide in addition the corresponding Fuller(4) (see Section A.8.2.1) and PULSE estimates. Since we do not have access to interventional data, we cannot directly test the distributional robustness properties discussed in Section 2.2.3. For the third study, however, the exogenous variable is continuous, which allows us to investigate distributional robustness empirically by holding out data points with extreme values of the exogenous variable and predict on these held-out data.

For the remainder of this section we use the PULSE estimator with $p_{\min} = 0.05$ and the test scaling-scheme that renders the test equivalent to the asymptotic version of the Anderson-Rubin test (see Section 2.3.2). Code replicating this analysis is available on GitHub.[1]

A.9.1. Angrist and Krueger (1991)

The dataset of Angrist and Krueger (1991) consists, in part, of 1980 US census data of 329,509 men born between 1930–1939. The endogenous target of interest is log weakly wages and the main endogenous regressor is years of education is instrumented on year and quarter of birth indicators. We consider four models M1–M4 corresponding to the models presented in column (1)–(8) in Table 5 of

[1] `https://github.com/MartinEmilJakobsen/PULSE/tree/master/Empirical_Applications`

Angrist and Krueger (1991). Model M1 is given by the structural reduced form equations

$$\log \text{weakly wage} = \text{educ} \cdot \gamma + \sum_i \text{YR}_i \cdot \beta_i + U_1,$$
$$\text{educ} = \sum_i \text{YR}_i \cdot \delta_i + \sum_{i,j} \text{YR}_i \cdot \text{QOB}_j \cdot \delta_{i,j} + U_2,$$

where educ is years of education, (YR_i) is year of birth indicators and (QOB_j) is quarter of birth indicators. Model M2 is given by M1 with the additional included exogenous regressors of age and age-squared. Models M3 and M4 are given by model M1 and M2, respectively, with additional included exogenous indicators describing race, marital status, metropolitan area and eight regional indicators. All models are over-identified, instrumenting education on a total of 30 binary instruments.

Table A.2 shows the OLS and TSLS estimates, as well as the Fuller(4) and PULSE estimates for the linear effect of education on log weakly wages. In all models the PULSE estimates coincide with the OLS estimates.

Table A.2: The estimated return of education on log weakly wage.

Model	OLS	TSLS	FUL	PULSE	Message	Test	Threshold
M1	0.0711	0.0891	0.0926	0.0711	OLS Acc.	26.92	55.76
M2	0.0711	0.0760	0.0739	0.0711	OLS Acc.	23.15	55.76
M3	0.0632	0.0806	0.0835	0.0632	OLS Acc.	23.79	68.67
M4	0.0632	0.0600	0.0555	0.0632	OLS Acc.	19.59	68.67

Note: Point estimates for the return of education on log weakly wage. The OLS and TSLS values coincide with the ones in Table V of Angrist and Krueger (1991). The right columns show the values of the test statistic (evaluated in the PULSE estimates) and the test rejection thresholds. For all models, the OLS is accepted and the PULSE coincides with the OLS.

A.9.2. Card (1993)

The dataset of Card (1993) consists of a US National Longitudinal Survey of Young Men spanning from 1966 to 1981. The subset of interest consists of 3010 observations for which there is recorded a valid wage and education level in a 1976 interview. The endogenous target of interest is log hourly wages and the main endogenous regressor is years of education. Proximity to a four year college, recorded in 1966, is used as an instrument. We consider two models, M1 and M2, corresponding to models in Panel B, column (5) and (6) of Table 3 (Card, 1993), respectively. Model M1 is given by regressing the target, log hourly wages, on included exogenous indicators of race, metropolitan area and region; the included endogenous regressors are years of education, work-experience and work-experience-squared. The endogenous regressors are instrumented by the excluded exogenous

variables age, age-squared and indicator of proximity to college. In model M2, we have model M1 with the addition of several exogenous indicators of parents education level.

Table A.3 shows the OLS and TSLS estimates, as well as the Fuller(4) and PULSE estimates for the linear effect of education on log hourly wages. Again, in all models the OLS estimates are not rejected by the Anderson-Rubin test. Hence, all PULSE estimates coincide with the OLS estimates.

Table A.3: The estimated return of education on log hourly wages.

Model	OLS	TSLS	FUL	PULSE	Message	Test	Thresh.
M1	0.0747	0.1224	0.1156	0.0747	OLS Acc.	1.22	26.30
M2	0.0726	0.1324	0.1283	0.0726	OLS Acc.	1.71	43.77

Note: Point estimates for the return of education on log hourly wage. The OLS and TSLS values coincide with the ones shown in Table 3 of Card (1993). The right columns show the values of the test statistic (evaluated in the PULSE estimates) and the test rejection thresholds. For all models, the OLS is accepted and the PULSE coincides with the OLS.

A.9.3. Acemoglu et al. (2001)

In Section 2.5.1 of the main paper we describe the data and models of Acemoglu et al. (2001). Furthermore, we replicate the OLS and TSLS estimates and presented the corresponding Fuller(4) and PULSE estimates.

To investigate distributional robustness, we conduct an out-of-sample mean squared prediction error (MSPE) analysis on a mean-centered dataset of the just-identified identified model M1. This is the simplest model proposed in Acemoglu et al. (2001) but the MSPE robustness property of Theorem 2.1 is robust to model misspecifications; see Remark A.1. We do not have access to interventional data. Instead, for different values of $n_{\text{test}} \in \mathbb{N}$, that is, for each $n_{\text{test}} \in \{4, 8, ..., 32\}$, we remove the data points with the $n_{\text{test}}/2$ lowest and $n_{\text{test}}/2$ highest settler mortality rates. We then fit the OLS, TSLS, PULSE and Fuller(4) on the remaining $64 - n_{\text{test}}$ observations and compute the out-of-sample MSPE on the n_{test} held-out observations, measuring the model's ability to generalize.

The instrument has a larger variance on the held-out data and the population robustness property of K-class estimators (see Theorem 2.1) suggests that PULSE and Fuller(4) might generalize slightly better than OLS or TSLS.[2] The results of this analysis is summarised in Table A.5. Indeed, we see that the OLS is optimal for a small number of held-out data points (when little generalization is required) and that for an increasing number of held-out data points, PULSE and FULLER(4) outperform the other estimators in terms of MSPE.

[2]Here, we consider a just-identified model, so the Fuller(4) K-class parameter $\kappa \in (0, 1)$.

For comparison, we also consider random sample splits, i.e., taking out a random subset of the dataset rather. Here, no generalization is required and as expected, OLS performs better than the other estimates, see Table A.4. The MSPE is minimized by OLS, PULSE, Fuller(4), and TSLS in 65.9%, 21.8%, 6.1%, and 6.2% of the cases, respectively.

Table A.4: log GPD MSPE orderings on random sample splits.

MSPE	Outperforms			
	OLS	PULSE	FUL	TSLS
OLS	✗	65.9%	79.7%	85.3%
PULSE	34.1%	✗	87.7%	90.5%
FUL	20.3%	12.3%	✗	93.8%
TSLS	14.7%	9.5%	6.2%	✗

Note: The table shows generalization performance for different estimators on model M1 of Acemoglu et al. (2001). The data set is split randomly into a subset of 90% of the data (that is, 58 observations) and the MSPE for the OLS, PULSE, Fuller(4), and TSLS are calculated on the remaining 10% of the data. This procedure is repeated 1000 times. The table shows how often the estimators outperform each other. E.g., OLS has lower MSPE than TSLS in 85.3% of the cases. Here, no generalziation is needed and, as expected, the OLS performs best.

Table A.5: log GPD MSPE on extreme out-of-sample instrument observations.

	Estimated coefficient				K-class κ		MSPE			
n_{test}	OLS	TSLS	PULSE	FUL	PULSE	FUL	OLS	TSLS	PULSE	FUL
4	0.5015	1.1592	0.7852	0.9509	0.8286	0.9322	**0.2072**	2.0358	0.3211	0.8613
6	0.5113	0.9441	0.6590	0.8313	0.7075	0.9298	**0.8282**	1.5889	0.8692	1.2034
8	0.5017	0.9433	0.6287	0.8150	0.6781	0.9273	0.7800	1.5331	**0.7796**	1.0961
10	0.4978	0.8795	0.5810	0.7717	0.5733	0.9245	0.7018	1.0850	**0.6769**	0.8479
12	0.4901	0.8693	0.5390	0.7512	0.4407	0.9216	0.6605	1.0346	**0.6357**	0.7788
14	0.4748	0.8439	0.4748	0.7091	0.0000	0.9184	**0.6562**	0.8910	**0.6562**	0.6722
16	0.4581	0.7655	0.4581	0.6359	0.0000	0.9149	0.7290	0.7581	0.7290	**0.6573**
18	0.4247	0.6861	0.4247	0.5451	0.0000	0.9111	0.7476	**0.6263**	0.7476	**0.6263**
20	0.3883	0.8604	0.3883	0.6096	0.0000	0.9070	0.8886	0.8354	0.8886	**0.6632**
22	0.3789	0.8867	0.3789	0.6046	0.0000	0.9024	0.8285	0.8315	0.8285	**0.6072**
24	0.3784	0.7016	0.3784	0.5450	0.0000	0.8974	0.9152	**0.7251**	0.9152	0.7334
26	0.4156	0.8753	0.5240	0.6723	0.6682	0.8919	0.8794	1.0333	**0.7957**	0.8012
28	0.4155	0.7867	0.4676	0.6306	0.4789	0.8857	0.8340	0.8530	0.7880	**0.7468**
30	0.4016	0.8725	0.4710	0.6278	0.5754	0.8788	0.7989	0.9223	0.7370	**0.6991**
32	0.4087	0.9103	0.4893	0.6228	0.6344	0.8710	0.7823	0.9880	0.7225	**0.7016**

Note: The table shows generalization performance for different estimators on model M1 of Acemoglu et al. (2001). We remove the n_{test} observations with the most extreme values of settler mortality, fit OLS, TSLS, PULSE and Fuller(4) on the $64 - n_{\text{test}}$ samples, and compute the MSPE on the n_{test} held-out samples (four right-most columns). Indeed, in particular for larger values of n_{test}, where more generalization is needed, PULSE and Fuller(4) outperform OLS and TSLS in the majority of cases. The columns "Estimated coefficient" show the estimates for the linear effect of average expropriation risk on log GPD of each estimation method. The column "K-class κ" shows K-class κ parameters for both the PULSE and Fuller(4) estimates; EQ is computed according to $\text{Var(out-of-sample)} = \text{Var(in-sample)}/(1 - \kappa_{\text{EQ}})$.

A.10. Weak Instruments

There is a wide variety of attempts to quantify weakness of instruments, see e.g. Andrews et al. (2019) and Stock et al. (2002) for an overview. Heuristically, the presence of weak instruments in a instrumental variable setup refers to the notion that the causal effects of the instruments onto regressors are weak relative to the noise variance of the regressors. This strength of the instruments has direct effects on the finite sample behavior of instrumental variable estimators. For simplicity consider a mean zero collapsed causal structural model with no included exogenous variables entering the equation of interest, that is,

$$Y = \gamma^\intercal X + U_Y, \qquad X = \xi^\intercal A + U_X, \tag{A.35}$$

where $A \in \mathbb{R}^q$ are the collection of exogenous variables and the noise variables U_X and U_Y are possibly correlated. Let $\mathbf{A}, \mathbf{X}, \mathbf{Y}$ be a n-sample data matrices of i.i.d. realizations of the system in Equation (A.35). A key statistic used to quantify weakness of instruments is the concentration matrix given by $\mu_n = \Sigma_{U_X}^{-1/2} \xi^\intercal \mathbf{A}^\intercal \mathbf{A} \xi \Sigma_{U_X}^{-1/2}$, where Σ_{U_X} is the variance matrix of U_X. This statistic turns up in numerous different aspect of the finite sample properties of the two-stage least square estimator. Rothenberg (1984) argues that the one-dimensional analogue of μ_n under deterministic instruments and normal distributed noise variables directly influences the goodness of approximating a finite sample standardized two-stage least square estimator by its Gaussian asymptotic distribution. He argues that for large concentration parameters the Gaussian approximation is good. The concentration parameter can also be connected to approximate bias of the two-stage least squares estimator. Under assumptions similar to the above, Nagar (1959) showed that an approximate (to the order of $\mathcal{O}(n^{-1})$) finite sample bias of the two-stage least square estimator is inversely proportional to μ_n. Note that the concentration matrix μ_n is not observable, but may be approximated by $\hat{\mu}_n = \hat{\Sigma}_{U_X}^{-1/2} \mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X} \hat{\Sigma}_{U_X}^{-1/2}$, where $\hat{\Sigma}_{U_X} = \frac{1}{n-q} \mathbf{X}^\intercal P_{\mathbf{A}}^\perp \mathbf{X}$ is an estimator of the variance matrix of U_X and $P_{\mathbf{A}} \mathbf{X}$ is the ordinary least square prediction of $\mathbf{A}\xi$. Now define

$$G_n := \frac{\hat{\mu}_n}{q} = \frac{\hat{\Sigma}_{U_X}^{-1/2} \mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X} \hat{\Sigma}_{U_X}^{-1/2}}{q},$$

which can be seen as a multivariate first-stage F-statistic for testing the hypothesis $H_0 : \xi = 0$. That is, when $X \in \mathbb{R}$, then $G_n = \frac{n-q}{q} \frac{\mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X}}{\mathbf{X}^\intercal \mathbf{X} - \mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X}}$ is recognized as the F-test for testing H_0. Stock and Yogo (2002) propose to reject the hypothesis of a presence of weak instruments if the test-statistic $\lambda_{\min}(G_n)$, the smallest eigenvalue of G_n, is larger than a critical value that, for example, depends on how much bias you allow your estimator to have. Prior to this G_n had been used to test under-identifiability in the sense that the concentration matrix is singular (Cragg and Donald, 1993), while the former uses a small minimum eigenvalue of G_n as a proxy for the presence of weak instruments in identified models. From the work

of Staiger and Stock (1997) a frequently appearing rule of thumb for instruments being non-weak is that the F-statistic G_n ($\lambda_{\min}(G_n)$ in higher dimensions) is larger than 10. A more formal justification of this rule is due to Stock and Yogo (2002) who showed (under weak-instrument asymptotics) that it approximately (in several models) corresponds to a 5% significance test that the bias of TSLS is at most 10% of the bias of OLS.

We can, under further model simplification, strengthen the intuition on how the concentration matrix G_n and especially the minimum eigenvalue $\lambda_{\min}(G_n)$ governs the weakness of instruments. To this end assume that $\mathrm{Var}(U_X) = \Sigma_{U_X} = I$ and note that $\hat{\mu}_n$ is approximately proportional to the Hessian of the two-stage least squares objective function. That is, $\hat{\mu}_n \approx \Sigma_{U_X}^{-1/2} \mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X} \Sigma_{U_X}^{-1/2} = \mathbf{X}^\intercal \mathbf{A} (\mathbf{A}^\intercal \mathbf{A})^{-1} \mathbf{A}^\intercal \mathbf{X} \propto H(l_{\mathrm{IV}}^n)$. Hence, we have that $\lambda_{\min}(G_n)$ is approximately proportional to the curvature of two-stage least squares objective function in the direction of least curvature. Thus, if $\lambda_{\min}(G_n)$ is small, then, heuristically, the objective function l_{IV}^n has weak identification in the direction of the corresponding eigenvector. That is, changes to the point estimate of β away from the two-stage least square solution in this direction does not have a strong effect on the objective value. Finally, the weak instrument problem is a small sample problem. To this end note that $n^{-1} G_n = n^{-1} \hat{\Sigma}_{U_X}^{-1/2} \mathbf{X}^\intercal P_{\mathbf{A}} \mathbf{X} \hat{\Sigma}_{U_X}^{-1/2} \xrightarrow{P} \mathrm{Var}(U_X)^{-1/2} \xi^\intercal \mathrm{Var}(A)^{-1} \xi \mathrm{Var}(U_X)^{-1/2}$, hence by the continuity of the minimum eigenvalue operator, we have that $\lambda_{\min}(G_n) \xrightarrow{P} \infty$.

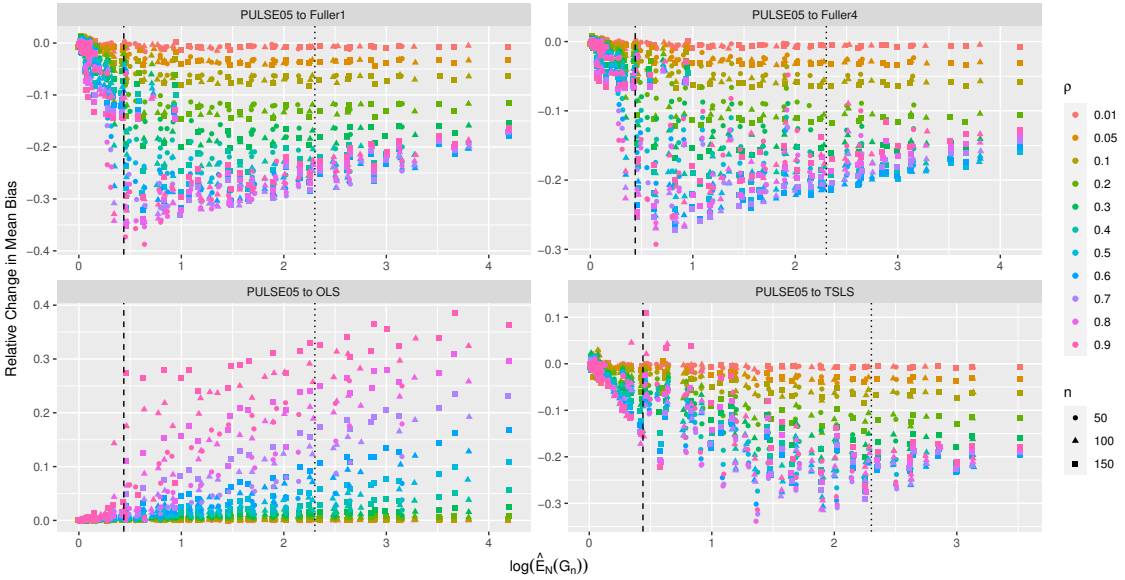


Figure A.8: Illustrations of the relative change in the absolute value of the mean bias (a positive relative change means that PULSE is better). The vertical dotted line corresponds to the rule of thumb for classifying instruments as weak, i.e., an F-test rejection threshold of 10. The first stage F-test for $H_0 : \bar{\xi} = 0$, i.e., for the relevancy of instruments, at a significance level of 5%, has different rejection thresholds in the range $[1.55, 4.04]$ depending on n and q. The vertical dashed line corresponds to the smallest rejection threshold of 1.55. Note that the lowest possible negative relative change is -1. For the comparison with the TSLS estimator we have removed the case $q = 1$ to ensure existence of first moments. TSLS, Fuller(1) and Fuller(4) outperforms PULSE while PULSE outperforms OLS.

A.11. Additional Simulation Experiments

A.11.1. Additional Illustrations for the Univariate Experiment

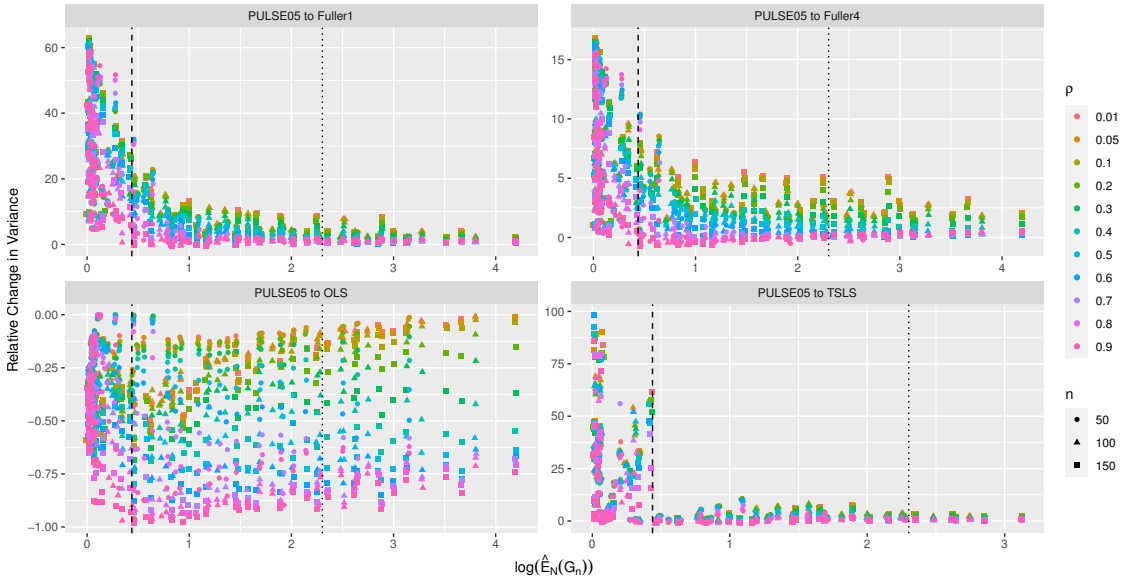


Figure A.9: Illustrations of the relative change in variance (a positive relative change means that PULSE is better). The vertical lines are identical to those of Figure A.8. For the comparison with the TSLS estimator we have removed the case $q \in \{1, 2\}$ to ensure existence of second moments. We have removed two observations with relative change above 100, in the very weak instrument setting, for aesthetic reasons. PULSE outperforms TSLS, Fuller(1) and Fuller(4), especially for low confounding and weak instruments. We also see that OLS outperforms PULSE with the largest decrease in variance for the large confounding cases.

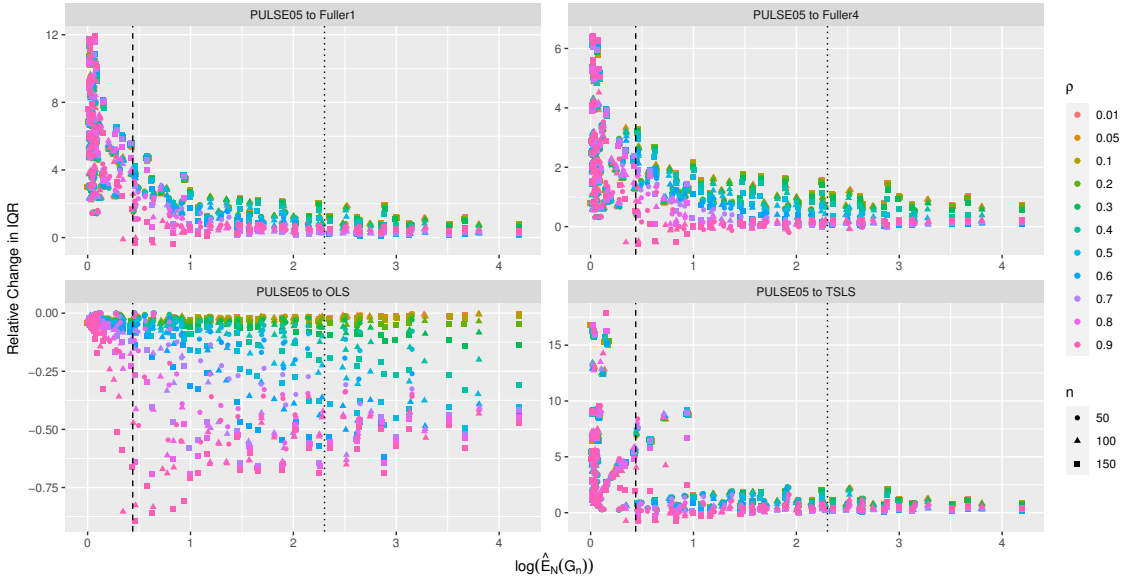


Figure A.10: Illustrations of the relative change in interquartile range (a positive relative change means that PULSE is better). The vertical lines are identical to those of Figure A.8. We see that PULSE is superior to Fuller(1), Fuller(4) and TSLS except in very few cases with very large confounding. Furthermore, OLS outperforms PULSE with relatively small difference for low confounding and larger difference for large confounding.

A.11.2. Additional Illustrations for the Multivariate Experiment

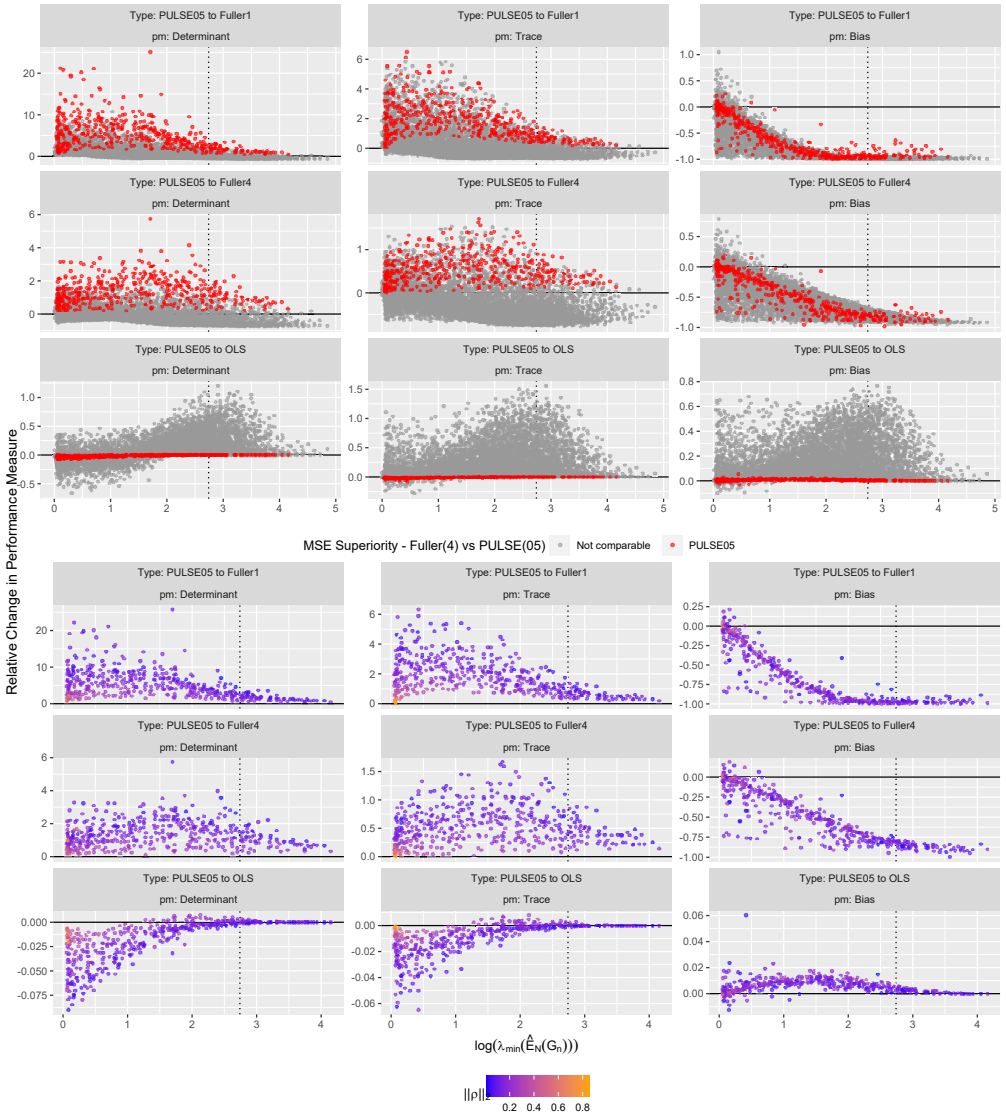


Figure A.11: There are two illustrations, both illustrating relative changes in performance measures as in Figure A.4 except that the points are color-graded according to MSE superiority when comparing Fuller(4) and PULSE (top 3×3) and confounding strength $\|\rho\|_2$ (bottom 3×3) . Among the 10000 randomly generated models there are 461 models where PULSE is MSE superior to Fuller(4). In the remaining 9539 models the MSE matrices are not comparable. For the 461 models where PULSE was MSE superior the simulations were repeated with $N = 25000$ repetitions to account for possible selection bias. Of the 461 models 445 were still superior when increasing N from 5000 to 25000. The bottom 3×3 grid is an illustration of the relative change in performance measure for the 445 models that remained superior, each model color-graded according to confounding strength. We see that in almost all of these models there is weak to moderate confounding. The exception being a few models in the very weak instrument setting where the confounding is strong.

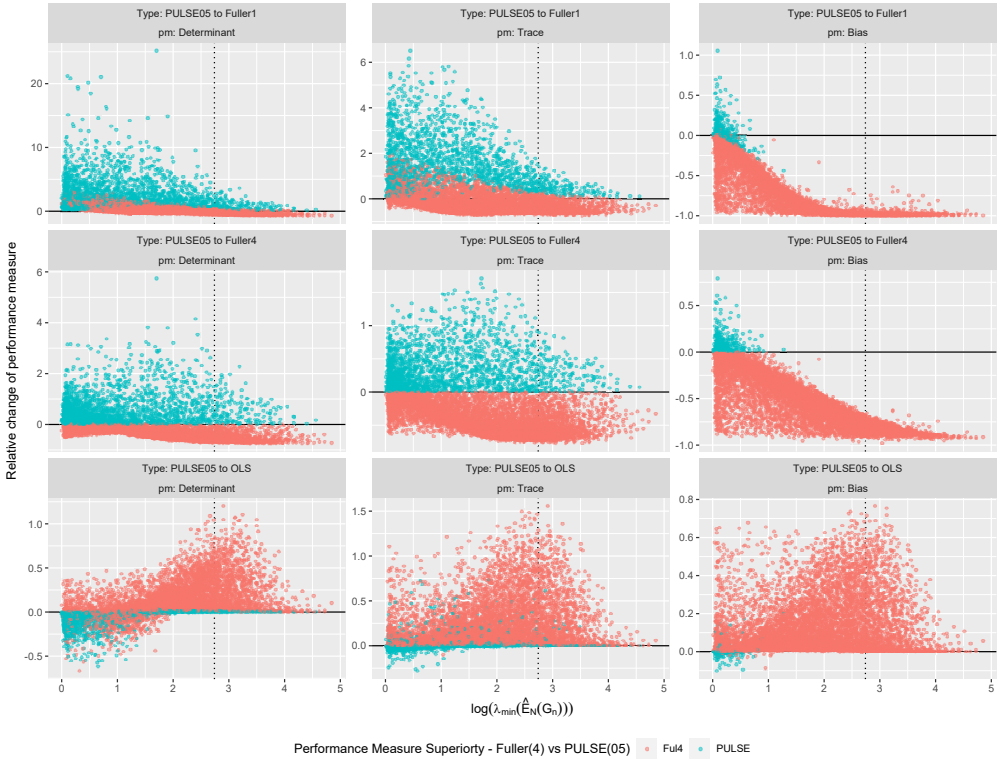


Figure A.12: This figure shows the same results as in Figure A.4 except that the points are color-graded according to performance measure superiority when comparing Fuller(4) and PULSE(05). That is, the models have fixed column-wise color-grading according to the comparison between Fuller(4) and PULSE(05).

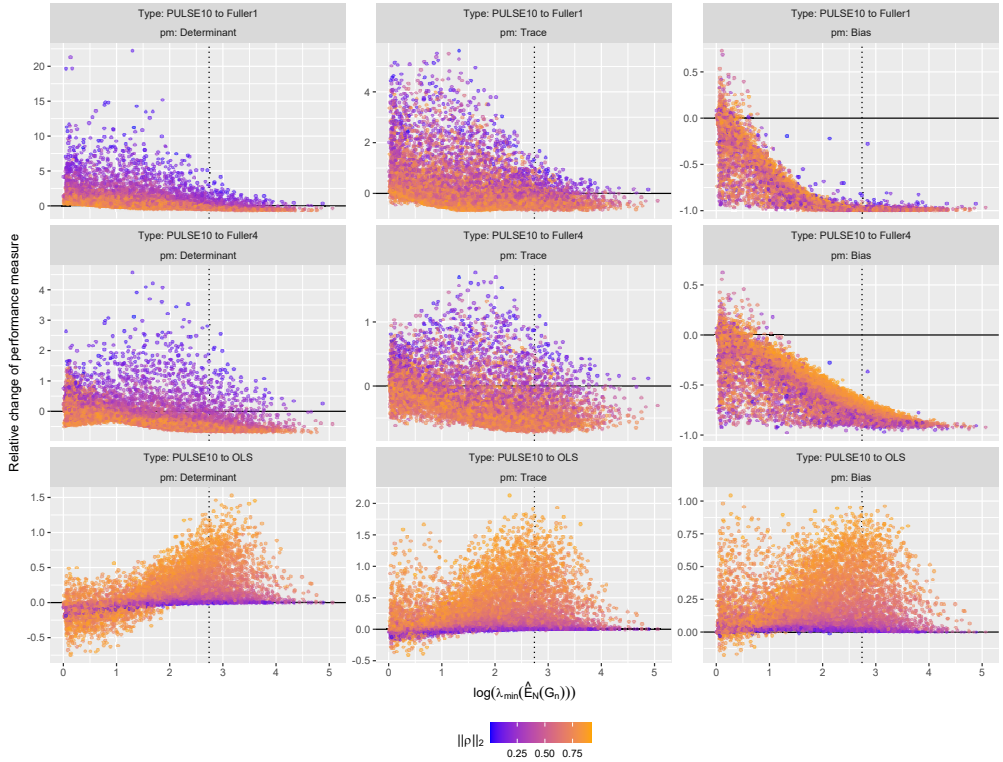


Figure A.13: This figure shows the same results as in Figure A.4 except that we here compare PULSE with $p_{\min} = 0.1$ to the benchmark estimators.

Appendix B

A Causal Framework for Distribution Generalization

B.1. Transforming Causal Models

As illustrated in Remark 3.1, our framework can also be applied in situations where training and test distributions are generated from an SCM with a different structure than (3.1). Below, we show that a general class of SCMs can be transformed into our reduced setting. To this end, assume the true underlying causal structure is given by the SCM

$$\begin{aligned} A &\coloneqq \varepsilon_A & X &\coloneqq w(X,Y) + g(A) + h_2(H, \varepsilon_X) \\ H &\coloneqq \varepsilon_H & Y &\coloneqq f(X) + h_1(H, \varepsilon_Y), \end{aligned} \tag{B.1}$$

where, as before, f, g, w, h_1 and h_2 are measurable functions. First, we show how to transform the above SCM into the reduced form (3.1) without changing the induced observational distribution. In Appendix B.1.1, we then discuss how to transform interventions in (B.1) to interventions in the reduced model.

Throughout this appendix, we assume that (B.1) is uniquely solvable in the sense that there exists a unique function F such that $(A, H, X, Y) = F(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ almost surely, see Bongers et al. (2021) for more details. Denote by F_X the coordinates of F that correspond to the X variable (i.e., the coordinates from $r+q+1$ to $r+q+d$). We further assume that there exist functions $\tilde{g}$ and $\tilde{h}_2$ such that

$$F_X(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y) = \tilde{g}(\varepsilon_A) + \tilde{h}_2((\varepsilon_H, \varepsilon_Y), \varepsilon_X). \tag{B.2}$$

This decomposition is not always possible, but it exists in the following settings, for example: (i) *There are no A variables.* In these cases, the additive decomposition

(B.2) becomes trivial. (ii) *There are further constraints on the original SCM.* The additive decomposition (B.2) holds if, for example, w is a linear function or A only enters the structural assignments of covariates X which have at most Y as a descendant.

Using the decomposition in (B.2), we can define the following reduced SCM

$$\begin{aligned} A &:= \varepsilon_A & X &:= \tilde{g}(A) + \tilde{h}_2(\tilde{H}, \varepsilon_X) \\ \tilde{H} &:= \varepsilon_{\tilde{H}} & Y &:= f(X) + h_1(\tilde{H}), \end{aligned} \tag{B.3}$$

where $\varepsilon_{\tilde{H}}$ has the same distribution as $(\varepsilon_H, \varepsilon_Y)$ in (B.1). This model fits the framework from Section 3.2.1, where the noise term in Y is now taken to be constantly zero. Both SCMs (B.1) and (B.3) induce the same observational distribution and the same function f appears in the assignments of Y.

If one intends to use interventions in the original SCM (i.e., (B.1)) to model the test distributions, one needs to also transform these interventions. We discuss how this can be done in the following subsection.

B.1.1. Transforming Interventions

For SCMs of the form (B.1) (and which satisfy (B.2)), any distribution arising from an intervention on a subset of covariates from X can be equivalently expressed using an intervention on all of X in the corresponding reduced model (B.3). To see this, let i be such an intervention in the original SCM, and let $\mathbb{P}^i$ be the induced interventional distribution over (X, Y, A). We can then generate the same intervention distribution in (B.3) using the intervention $X := \varepsilon_X^i$, where the distribution of ε_X^i coincides with the marginal of X in $\mathbb{P}^i$. Note, however, that this type of transformation may fail for some model classes, for example, this may happen if the original SCM contains a hidden variable which is a descendant of some (intervened) X variables and a cause of Y. Also, even in situations where the above transform is possible, the interventions can change their intervention targets, become non-well-behaved or change their support. In order to apply the developed methodology, one needs to check whether the transformed interventions are a well-behaved (this is not necessarily the case, even if the original intervention was well-behaved) and how the support of all X variables behaves under that specific intervention.

Intervention type First, we consider which types of interventions in (B.1) translate to well-behaved interventions in (B.3). A simple example is given by interventions on A in the original SCM, which result in the same interventions on A also in the reduced SCM. Similarly, performing hard interventions on all components of X in the original SCM leads to the same intervention in the reduced SCM, which is in particular both confounding-removing and confounding-preserving. For interventions on subsets of the X, this is not always the case. To see that, consider the following example

$$\begin{aligned} A &\coloneqq \varepsilon_A \\ X_1 &\coloneqq \varepsilon_1 \\ X_2 &\coloneqq Y + \varepsilon_2 \\ Y &\coloneqq X_1 + \varepsilon_Y \end{aligned} \quad \xrightarrow{\text{transform}} \quad \begin{aligned} A &\coloneqq \varepsilon_A \\ H &\coloneqq \varepsilon_Y \\ X &\coloneqq (\varepsilon_1, H + \varepsilon_1 + \varepsilon_2) \\ Y &\coloneqq X_1 + H, \end{aligned}$$

with $\varepsilon_A, \varepsilon_1, \varepsilon_2, \varepsilon_Y$ i.i.d. noise innovations. Here, the left hand side represents the original SCM and the right hand side corresponds to the reduced SCM fitting in our framework. Consider now, in the original SCM, the intervention $X_1 \coloneqq i$, for some $i \in \mathbb{R}$. In the reduced SCM, this intervention corresponds to the intervention $X = (X_1, X_2) \coloneqq (i, H + i + \varepsilon_2)$, which is neither confounding-preserving nor confounding-removing.[1] On the other hand, any intervention on X_2 or A in the original SCM model corresponds to the same intervention in the reduced SCM. We can generalize these observations to the following statements

- *Interventions on A:* If we intervene on A in the original SCM (B.1) (i.e., by replacing the structural assignment of A with $\psi^i(I^i, \varepsilon_A^i)$), then this translates to the same intervention on A in the reduced SCM (B.3).
- *Shift intervention on X_j which are not ancestors of Y:* If we perform a shift intervention on X_j in the original SCM (B.1) (assuming no confounding H) and X_j is not an ancestor of Y, then this corresponds to a confounding-preserving intervention in the reduced SCM (B.3).
- *Hard interventions on all X:* If we intervene on all X in the original SCM (B.1) by replacing the structural assignment of X with an independent random variable $I \in \mathbb{R}^d$, then this translates to the same intervention in the reduced SCM (B.3) which is confounding-removing.
- *No X is a descendant of Y and there is no unobserved confounding H:* If we intervene on X in the original SCM (B.1) (i.e., by replacing the structural assignment of X with $\psi^i(g, A^i, \varepsilon_X^i, I^i)$), then this translates to a potentially different but confounding-removing intervention in the reduced SCM (B.3). This is because the reduced SCM (B.3) does not include unobserved variables H in this case.
- *Hard interventions on a variable X_j which has at most Y as a descendant:* If we intervene on X_j in the original SCM (B.1) by replacing the structural assignment of X_j with an independent random variable I, then this intervention translates to a potentially different but confounding-preserving intervention.

Other settings may yield well-behaved interventions, too, but may require more assumptions on the original SCM model (B.1) or further restrictions on the intervention classes.

[1] This may not come as a surprise since, without the help of an instrument, it is impossible to distinguish whether a covariate is an ancestor or a descendant of Y.

Intervention support A support-reducing intervention in the original SCM can translate to a support-extending intervention in the reduced SCM. Consider the following example

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= X_1 + \mathbf{1}\{X_1 = 0.5\} \\ Y &:= X_2 + \varepsilon_Y \end{aligned} \quad \xrightarrow{\text{transform}} \quad \begin{aligned} X &:= (\varepsilon_1, \varepsilon_1 + \mathbf{1}\{\varepsilon_1 = 0.5\}) \\ Y &:= X_2 + \varepsilon_Y, \end{aligned}$$

with $\varepsilon_1, \varepsilon_Y \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. As before, the left hand side represents the original SCM, whereas the right hand side corresponds to the reduced SCM converted to fit our framework. Under the observational distribution, the support of X_1 and X_2 is equal to the open interval $(0,1)$. Consider now the support-reducing intervention $X_1 := 0.5$ in original SCM. Within our framework, such an intervention would correspond to the intervention $X = (X_1, X_2) := (0.5, 1.5)$, which is support-extending. This example is rather special in that the SCM consists of a function that changes on a null set of the observational distribution. With appropriate assumptions to exclude similar degenerate cases, it is possible to show that support-reducing interventions in (B.1) correspond to support-reducing interventions within our framework (B.3).

B.2. Sufficient Conditions for Assumption 1 in IV Settings

Assumption 3.1 states that f is identified on the support of X from the observational distribution of (Y, X, A). Whether this assumption is satisfied depends on the structure of $\mathcal{F}$ but also on the other function classes $\mathcal{G}, \mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{Q}$ that make up the model class $\mathcal{M}$ from which we assume that the distribution of (Y, X, A) is generated.

Identifiability of the causal function in the presence of instrumental variables is a well-studied problem in econometrics literature. Most prominent is the literature on identification in linear SCMs (e.g., Fisher, 1966; Greene, 2003). However, identification has also been studied for various other parametric function classes. We say that $\mathcal{F}$ is a parametric function class if it can be parametrized by some finite dimensional parameter set $\Theta \subseteq \mathbb{R}^p$. We here consider classes of the form

$$\mathcal{F} := \{f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R} \,|\, \Theta \ni \theta \mapsto f(x, \theta) \in C^2, \forall x \in \mathbb{R}^d\}.$$

Consistent estimation of the parameter θ_0 using instrumental variables in such function classes has been studied extensively in the econometric literature (e.g., Amemiya, 1974; Jorgenson and Laffont, 1974; Kelejian, 1971). These works also contain rigorous results on how instrumental variable estimators of θ_0 are constructed and under which conditions consistency (and thus identifiability) holds. Here, we give an argument on why the presence of the exogenous variables A yields

identifiability under certain regularity conditions. Assume that $\mathbb{E}[h_1(H, \varepsilon_Y)|A] = 0$, which implies that the true causal function $f(\cdot, \theta_0)$ satisfies the population orthogonality condition

$$\mathbb{E}[l(A)^\top (Y - f(X, \theta_0))] = \mathbb{E}\Big[l(A)^\top \mathbb{E}[h_1(H, \varepsilon_Y)|A]\Big] = 0, \tag{B.4}$$

for some measurable mapping $l : \mathbb{R}^q \to \mathbb{R}^g$, for some $g \in \mathbb{N}_{>0}$. Clearly, θ_0 is identified from the observational distribution if the map $\theta \mapsto \mathbb{E}[l(A)^\top (Y - f(X, \theta))]$ is zero if and only if $\theta = \theta_0$. Furthermore, since $\theta \mapsto f(x, \theta)$ is differentiable for all $x \in \mathbb{R}^d$, the mean value theorem yields that, for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, there exists an intermediate point $\tilde{\theta}(x, \theta, \theta_0)$ on the line segment between θ and θ_0 such that

$$f(x, \theta) - f(x, \theta_0) = D_\theta f(x, \tilde{\theta}(x, \theta, \theta_0))(\theta - \theta_0),$$

where, for each $x \in \mathbb{R}^d$, $D_\theta f(x, \theta) \in \mathbb{R}^{1 \times p}$ is the derivative of $\theta \mapsto f(x, \theta)$ evaluated in θ. Composing the above expression with the random vector X, multiplying with $l(A)$ and taking expectations yields that

$$\begin{aligned} &\mathbb{E}[l(A)(Y - f(X, \theta_0))] - \mathbb{E}[l(A)(Y - f(X, \theta))] \\ &= \mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))](\theta_0 - \theta). \end{aligned}$$

Hence, if $\mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))] \in \mathbb{R}^{g \times p}$ is of rank p for all $\theta \in \Theta$ (which implies $g \geq p$), then θ_0 is identifiable as it is the only parameter that satisfies the population orthogonality condition of (B.4). As θ_0 uniquely determines the entire function, we get identifiability of $f \equiv f(\cdot, \theta_0)$, not only on the support of X but the entire domain $\mathbb{R}^d$, i.e., both Assumptions 3.1 and 3.2 are satisfied. In the case that $\theta \mapsto f(x, \theta)$ is linear, i.e. $f(x, \theta) = f(x)^T \theta$ for all $x \in \mathbb{R}^d$, the above rank condition reduces to $\mathbb{E}[l(A) f(X)^T] \in \mathbb{R}^{g \times p}$ having rank p (again, implying that $g \geq p$). Furthermore, when $(x, \theta) \mapsto f(x, \theta)$ is bilinear, a reparametrization of the parameter space ensures that $f(x, \theta) = x^T \theta$ for $\theta \in \Theta \subseteq \mathbb{R}^d$. In this case, the rank condition can be reduced to the well-known rank condition for identification in a linear SCM, namely that $\mathbb{E}[AX^T] \in \mathbb{R}^{q \times p}$ is of rank p.

Finally, identifiability and methods of consistent estimation of the causal function have also been studied for non-parametric function classes. The conditions for identification are rather technical, however, and we refer the reader to Newey (2013); Newey and Powell (2003) for further details.

B.3. Choice of Test Statistic

By considering the variables

$$B(X) = (B_1(X), \ldots, B_k(X)) \text{ and } C(A) = (C_1(A), \ldots, C_k(A)),$$

as vectors of covariates and instruments, respectively, our setting in Section 3.5.2 reduces to the classical (just-identified) linear IV setting. We could therefore use

a test statistics similar to the one propsed by the PULSE (Jakobsen and Peters, 2021). With a notation that is slightly adapted to our setting, this estimator tests $\tilde{H}_0(\theta)$ using the test statistic

$$T_n^1(\theta) = c(n)\frac{\|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2}{\|\mathbf{Y} - \mathbf{B}\theta\|_2^2},$$

where $\mathbf{P}$ is the projection onto the columns of $\mathbf{C}$, and $c(n)$ is some function with $c(n) \sim n$ as $n \to \infty$. Under the null hypothesis, T_n^1 converges in distribution to the χ_k^2 distribution, and diverges to infinity in probability under the general alternative. Using this test statistic, $\tilde{H}_0(\theta)$ is rejected if and only if $T_n^1(\theta) > q(\alpha)$, where $q(\alpha)$ is the $(1-\alpha)$-quantile of the χ_k^2 distribution. The acceptance region of this test statistic is asymptotically equivalent with the confidence region of the Anderson-Rubin test Anderson and Rubin (1949) for the causal parameter θ^0. Using the above test results in a consistent estimator for θ^0 (Jakobsen and Peters, 2021, Theorem 3.12); the proof exploits the particular form of T_n^1 without explicitly imposing that assumptions (C1) and (C2) hold.

If the number k of basis functions is large, however, numerical experiments suggest that the above test has low power in finite sample settings. As default, our algorithm therefore uses a different test based on a penalized regression approach. This test has been proposed in Chen et al. (2014) for inference in nonparametric regression models. We now introduce this procedure with a notation that is adapted to our setting. For every $\theta \in \mathbb{R}^k$, let $R_\theta = Y - B(X)^\top\theta$ be the residual associated with θ. We then test the slightly stronger hypothesis

$$\bar{H}_0(\theta) : \exists\, \sigma_\theta^2 > 0 \text{ s.t. } \mathbb{E}[R_\theta \,|\, A] \overset{\text{a.s}}{=} 0 \text{ and } \mathrm{Var}[R_\theta \,|\, A] = \sigma_\theta^2$$

against the alternative that $\mathbb{E}[R_\theta \,|\, A] = m(A)$ for some smooth function m. To see that the above hypothesis implies $\tilde{H}_0(\theta)$ (and therefore $H_0(\theta)$, see Section 3.5.2.1), let $\theta \in \mathbb{R}^k$ be such that $\bar{H}_0(\theta)$ holds true. Then,

$$\begin{aligned}
\mathbb{E}[C(A)(Y - B(X)^\top\theta)] &= \mathbb{E}[C(A)R_\theta] \\
&= \mathbb{E}[\mathbb{E}[C(A)R_\theta \,|\, A]] \\
&= \mathbb{E}[C(A)\mathbb{E}[R_\theta \,|\, A]] \\
&= 0,
\end{aligned}$$

showing that also $\tilde{H}_0(\theta)$ holds true. Thus, if $\tilde{H}_0(\theta)$ is false, then also $\bar{H}_0(\theta)$ is false. As a test statistic $T_n^2(\theta)$ for $\bar{H}_0(\theta)$, we use (up to a normalization) the squared norm of a penalized regression estimate of m, evaluated at the data $\mathbf{A}$, i.e., the TSLS loss $\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In the fixed design case, where $\mathbf{A}$ is non-random, it has been shown that, under $\bar{H}_0(\theta)$ and certain additional regularity conditions, it holds that

$$\frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \sigma_\theta^2 c_n}{\sigma_\theta^2 d_n} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where c_n and d_n are known functions of $\mathbf{C}$, $\mathbf{M}$ and δ (Chen et al., 2014, Theorem 1). The authors further state that the above convergence is unaffected by exchanging σ_θ^2 with a consistent estimator $\hat{\sigma}_\theta^2$, which motivates our use of the test statistic

$$T_n^2(\theta) := \frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \hat{\sigma}_{\theta,n}^2 c_n}{\hat{\sigma}_{\theta,n}^2 d_n},$$

where $\hat{\sigma}_{\theta,n}^2 := \frac{1}{n-1}\sum_{i=1}^n \|(\mathbf{I}_n - \mathbf{P}_\delta)(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. As a rejection threshold $q(\alpha)$ we use the $1-\alpha$ quantile of a standard normal distribution. For results on the asymptotic power of the test defined by T^2, we refer to Section 2.3 in Chen et al. (2014).

In our software package, both of the above tests are available options.

B.4. Addition to Experiments

B.4.1. Sampling of the Causal Function

To ensure linear extrapolation of the causal function, we have chosen a function class consisting of natural cubic splines, which, by construction, extrapolate linearly outside the boundary knots. We now describe in detail how we sample functions from this class for the experiments in Section 3.5.2.4. Let $q_{\min}$ and $q_{\max}$ be the respective 5%- and 95% quantiles of X, and let $B_1, \dots, B_4$ be a basis of natural cubic splines corresponding to 5 knots placed equidistantly between $q_{\min}$ and $q_{\max}$. We then sample coefficients $\beta_i \overset{\text{iid}}{\sim} \text{Uniform}(-1, 1)$, $i = 1, \dots, 4$, and construct f as $f = \sum_{i=1}^4 \beta_i B_i$. For illustration, we have included 18 realizations in Figure B.1.

B.4.2. Violations of the Linear Extrapolation Assumption

We have assumed that the true causal function extrapolates linearly outside the 90% quantile range of X. We now investigate the performance of our method for violations of this assumption. To do so, we again sample from the model (3.4), with $\alpha_A = \alpha_H = \alpha_\varepsilon = 1/\sqrt{3}$. For each data set, the causal function is sampled as follows. Let $q_{\min}$ and $q_{\max}$ be the 5%- and 95% quantiles of X. We first generate a function $\tilde{f}$ that linearly extrapolates outside $[q_{\min}, q_{\max}]$ as described in Section B.4.1. For a given threshold κ, we then draw $k_1, k_2 \overset{\text{iid}}{\sim} \text{Uniform}(-\kappa, \kappa)$ and construct f for every $x \in \mathbb{R}$ by

$$f(x) = \tilde{f}(x) + \frac{1}{2}k_1((x - q_{\min})_-)^2 + \frac{1}{2}k_2((x - q_{\max})_+)^2,$$

such that the curvature of f on $(-\infty, q_{\min}]$ and $[q_{\max}, \infty)$ is k_1 and k_2, respectively. Figure B.2 shows results for $\kappa = 0, 1, 2, 3, 4$. As the curvature increases, the ability to generalize decreases.

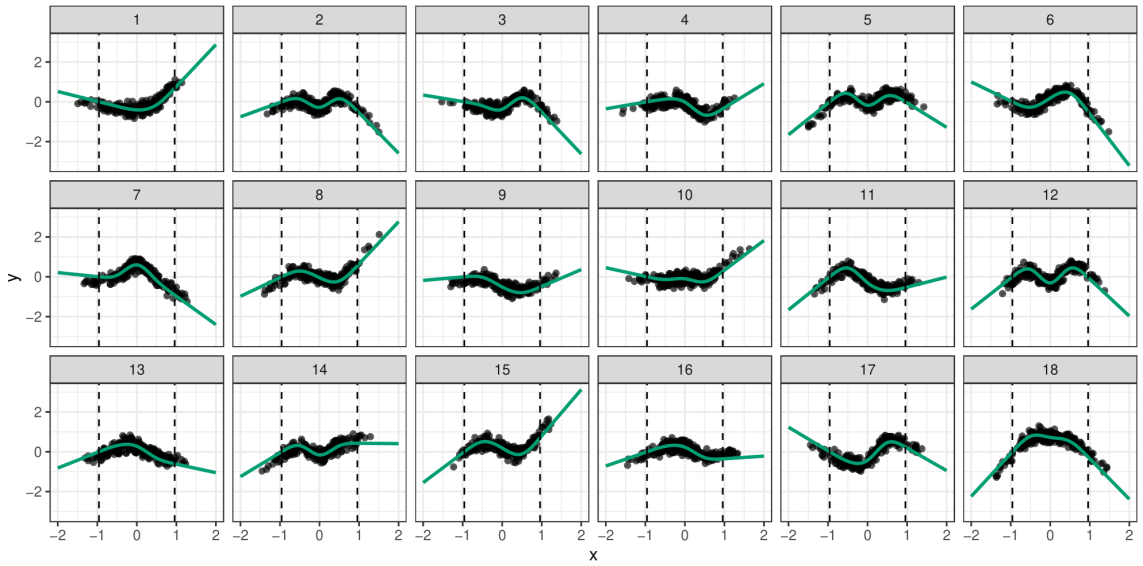


Figure B.1: The plots show independent realizations of the causal function that is used in all our experiments. These are sampled from a linear space of natural cubic splines, as described in Appendix B.4.1. To ensure a fair comparison with the alternative method, NPREGIV, the true causal function is chosen from a model class different from the one assumed by the NILE.

B.4.3. Running NILE on Half of the Available Data

In Section 3.5.2.4, we compared the NILE to several alternative procedures for estimating a non-linear causal function. As mentioned, these procedure use a sample-splitting strategy, where the two steps of the the two-stage-least-squares procedure are run on disjoint data sets. The NILE, on the other hand, uses all of the available data for the model fitting. Figure B.3 shows that, even when using only half of the available data, the NILE still outperforms the other methods considerably.

B.5. Proofs

Proof of Proposition 3.1: Assume that $\mathcal{I}$ is a set of interventions on X with at least one confounding-removing intervention. Let $i \in \mathcal{I}$ and $f_\diamond \in \mathcal{F}$, then we have the following expansion

$$\begin{aligned}\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] \\ + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))],\end{aligned} \tag{B.5}$$

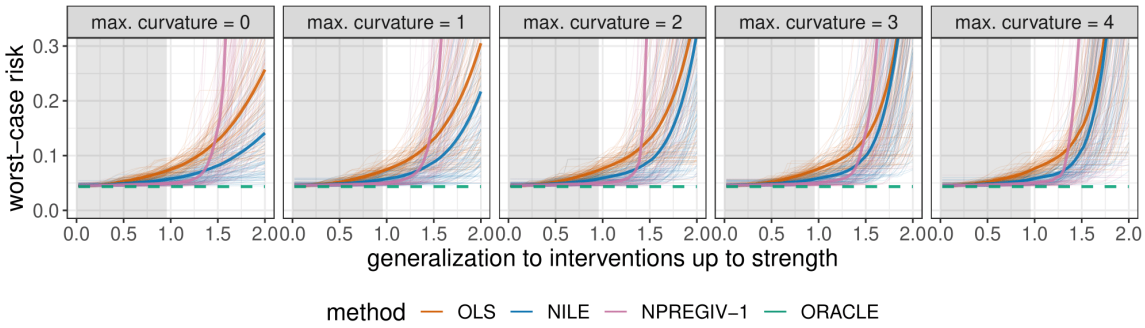


Figure B.2: Worst-case risk for increasingly strong violations of the linear extrapolation assumption. The grey area marks the inner 90 % quantile range of X in the training distribution. As the curvature of f outside the domain of the observed data increases, it becomes difficult to predict the interventional behavior of Y for strong interventions. However, even in situations where the linear extrapolation assumption is strongly violated, it remains beneficial to extrapolate linearly.

where $\xi_Y = h_1(H, \varepsilon_Y)$. For any intervention $i \in \mathcal{I}$ the causal function f always yields an identical loss. In particular, it holds that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2], \tag{B.6}$$

where we used that the distribution of ξ_Y is not affected by an intervention on X. The loss of the causal function can never be better than the minimax loss, that is,

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2]. \tag{B.7}$$

In other words, the minimax solution (if it exists) is always better than or equal to the causal function. We will now show that when $\mathcal{I}$ contains at least one confounding-removing intervention, then the minimax loss is dominated by any such intervention.

Fix $i_0 \in \mathcal{I}$ to be a confounding-removing intervention and let (X, Y, H, A) be generated by the SCM $M(i_0)$. Recall that there exists a map ψ^{i_0} such that $X := \psi^{i_0}(g, h_2, A, H, \varepsilon_X, I^{i_0})$ and that $X \perp\!\!\!\perp H$ as i_0 is a confounding-removing intervention. Furthermore, since the vectors A, H, ε_X, ε_Y and I^{i_0} are mutually independent, we have that $(X, H) \perp\!\!\!\perp \varepsilon_Y$ which together with $X \perp\!\!\!\perp H$ implies X, H and ε_Y are mutually independent, and hence $X \perp\!\!\!\perp h_1(H, \varepsilon_Y)$. Using this independence we get that $\mathbb{E}_{M(i_0)}[\xi_Y(f(X) - f_\diamond(X))] = \mathbb{E}_M[\xi_Y]\mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))]$. Hence, (B.5) for the intervention i_0 together with the modeling assumption $\mathbb{E}_M[\xi_Y] = 0$ implies that for all $f_\diamond \in \mathcal{F}$,

$$\mathbb{E}_M[\xi_Y^2] \leq \mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] = \mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2].$$

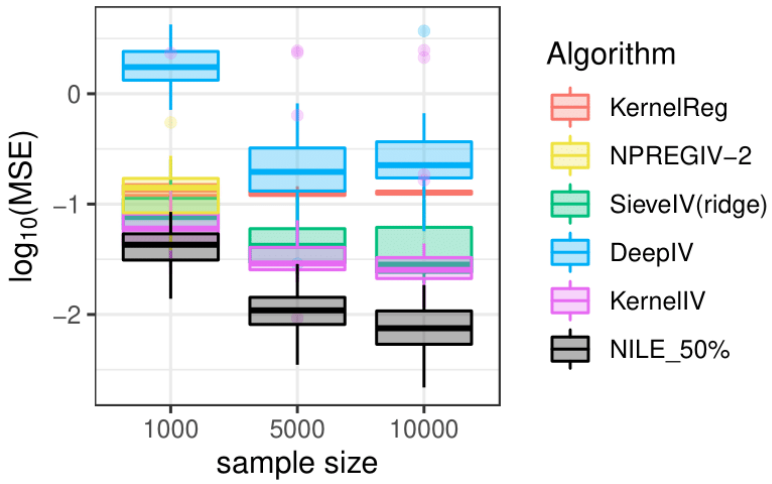


Figure B.3: Same results as shown in Figure 3.5, except that here, NILE is run only on half of the available data.

This proves that the smallest loss at a confounding-removing intervention is achieved by the causal function. Denoting the non-empty subset of confounding-removing interventions by $\mathcal{I}_{\text{cr}} \subseteq \mathcal{I}$, this implies

$$\begin{aligned}\mathbb{E}_M[\xi_Y^2] &= \inf_{f_\diamond \in \mathcal{F}} \mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2] \leq \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}_{\text{cr}}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \\ &\leq \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]. \end{aligned} \tag{B.8}$$

Combining (B.7) and (B.8) it immediately follows that

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

and hence

$$f \in \arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2],$$

which completes the proof of Proposition 3.1. □

Proof of Proposition 3.2: Let $\mathcal{F}$ be the class of all linear functions and let $\mathcal{I}$ denote the set of interventions on X that satisfy

$$\sup_{i \in \mathcal{I}} \lambda_{\min}\left(\mathbb{E}_{M(i)}\left[XX^\top\right]\right) = \infty.$$

We claim that the causal function $f(x) = b^\top x$ is the unique minimax solution of (3.2). We prove the result by contradiction. Let $\bar{f} \in \mathcal{F}$ (with $\bar{f}(x) = \bar{b}^\top x$) be such that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - \bar{b}^\top X)^2] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - b^\top X)^2],$$

and assume that $\|\bar{b} - b\|_2 > 0$. For a fixed $i \in \mathcal{I}$, we get the following bound

$$\begin{aligned}\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] &= (b - \bar{b})^\top \mathbb{E}_{M(i)}[XX^\top](b - \bar{b}) \\ &\geq \lambda_{\min}(\mathbb{E}_{M(i)}[XX^\top])\|b - \bar{b}\|_2^2.\end{aligned}$$

Since we assumed that the minimal eigenvalue is unbounded, this means that we can choose $i \in \mathcal{I}$ such that $\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2]$ can be arbitrarily large. However, applying Proposition 3.3, this leads to a contradiction since $\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] \leq 4\,\mathrm{Var}_M(\xi_Y)$ cannot be satisfied. Therefore, it must holds that $\bar{b} = b$, which moreover implies that f is indeed a solution to the minimax problem $\arg\min_{f_\diamond \in \mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]$, as it achieves the lowest possible objective value. This completes the proof of Proposition 3.2. □

Proof of Proposition 3.3: Let $\mathcal{I}$ be a set of interventions on X or A and let $f_\diamond \in \mathcal{F}$ with

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2]. \tag{B.9}$$

For any $i \in \mathcal{I}$, the Cauchy-Schwartz inequality implies that

$$\begin{aligned}\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] &= \mathbb{E}_{M(i)}[(f(X) + \xi_Y - f_\diamond(X))^2] \\ &= \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] \\ &\quad + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))] \\ &\geq \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \\ &\quad - 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}}.\end{aligned}$$

A similar computation shows that the causal function f satisfies

$$\mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2].$$

So by condition (B.9) this implies for any $i \in \mathcal{I}$ that

$$\begin{aligned}\mathbb{E}_M[\xi_Y^2] \geq\, &\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \\ &- 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}},\end{aligned}$$

which is equivalent to

$$\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 2\sqrt{\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]},$$

i.e. $\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4\mathbb{E}_M[\xi_Y^2]$. As this inequality holds for all $i \in \mathcal{I}$, we can take the supremum over all $i \in \mathcal{I}$, which completes the proof of Proposition 3.3. □

Proof of Proposition 3.4: As argued before, we have that for all $i \in \mathcal{I}_1$,

$$\mathbb{E}_{M(i)}\left[(Y - f(X))^2\right] = \mathbb{E}_{M(i)}\left[\xi_Y^2\right] = \mathbb{E}_M\left[\xi_Y^2\right].$$

Let now $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, using that the causal function f lies in $\mathcal{F}$, it holds that

$$\sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big] \leq \sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\Big[(Y - f(X))^2\Big] = \mathbb{E}_M\Big[\xi_Y^2\Big].$$

Moreover, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$, then it must also hold that

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big] \leq \mathbb{E}_M\Big[\xi_Y^2\Big] = \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\Big[(Y - f(X))^2\Big].$$

To prove the second part, we give a one-dimensional example. Let $\mathcal{F}$ be linear (i.e., $f(x) = bx$) and let $\mathcal{I}_1$ consist of shift interventions on X of the form

$$X^i \coloneqq g(A^i) + h_2(H^i, \varepsilon_X^i) + c,$$

with $c \in [0, K]$. Then, the minimax solution f_1^* (where $f_1^*(x) = b_1^* x$) with respect to $\mathcal{I}_1$ is not equal to the causal function f as long as $\mathrm{Cov}(X, \xi_Y)$ is strictly positive. This can be seen by explicitly computing the OLS estimator for a fixed shift c and observing that the worst-case risk is attained at $c = K$. Now let $\mathcal{I}_2$ be a set of interventions of the same form as $\mathcal{I}_1$ but including shifts with $c > K$ such that $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$. Since $\mathcal{F}$ consists of linear functions, we know that the loss $\mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big]$ can become arbitrarily large, since

$$\begin{aligned}
&\mathbb{E}_{M(i)}\Big[(Y - f_1^*(X))^2\Big] \\
&= (b - b_1^*)^2 \mathbb{E}_{M(i)}[X^2] + \mathbb{E}_M[\xi_Y^2] + 2(b - b_1^*)\mathbb{E}_{M(i)}[\xi_Y X] \\
&= (b - b_1^*)^2 (c^2 + \mathbb{E}_M[X^2] + 2c\mathbb{E}_M[X]) + \mathbb{E}_M[\xi_Y^2] \\
&\quad + 2(b - b_1^*)(\mathbb{E}_M[\xi_Y X] + \mathbb{E}_M[\xi_Y]c),
\end{aligned}$$

and $(b - b^*)^2 > 0$. In contrast, the loss for the causal function is always $\mathbb{E}_M[\xi_Y^2]$, so the worst-case risk of f_1^* becomes arbitrarily worse than that of f. This completes the proof of Proposition 3.4. □

Proof of Proposition 3.5: Let $\varepsilon > 0$. By definition of the infimum, we can find $f^* \in \mathcal{F}$ such that

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f^*(X))^2\Big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big] \right| \leq \varepsilon.$$

Let now $\tilde{M} \in \mathcal{M}$ be s.t. $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By assumption, the left-hand side of the above inequality is unaffected by substituting M for $\tilde{M}$, and the result thus follows. □

Proof of Proposition 3.6:

We first show that the causal parameter β is not a minimax solution. Let $u := \sup \mathcal{I} < \infty$, since $\mathcal{I}$ is bounded, and take $b = \beta + 1/(\sigma u)$. By an explicit

computation we get that

$$\begin{aligned}\inf_{b_\diamond\in\mathbb{R}}\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-b_\diamond X)^2\Big] &\leq \sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-bX)^2\Big]\\ &=\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(\varepsilon_Y+\tfrac{1}{\sigma}H-\tfrac{1}{\sigma u}iH)^2\Big]\\ &=\sup_{i\in\mathcal{I}}\Big[1+\Big(1-\tfrac{i}{u}\Big)^2\Big]\\ &<2\\ &=\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-\beta X)^2\Big],\end{aligned}$$

where the last inequality holds because $0<1+(1-i/u)^2<2$ for all $i\in\mathcal{I}$, and since $\mathcal{I}\subseteq\mathbb{R}_{>0}$ is compact with upper bound u. Hence,

$$\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-\beta X)^2\Big]-\inf_{b_\diamond\in\mathbb{R}}\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-b_\diamond X)^2\Big]>0,$$

proving that the causal parameter is not a minimax solution for model M w.r.t. $(\mathcal{F},\mathcal{I})$. Recall that in order to prove that $(\mathbb{P}_M,\mathcal{M})$ does not generalize with respect to $\mathcal{I}$ we have to show that there exists an $\varepsilon>0$ such that for all $b\in\mathbb{R}$ it holds that

$$\sup_{\tilde{M}:\mathbb{P}_{\tilde{M}}=\mathbb{P}_M}\Big|\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\Big[(Y-bX)^2\Big]-\inf_{b_\diamond\in\mathbb{R}}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\Big[(Y-b_\diamond X)^2\Big]\Big|\geq\varepsilon.$$

Thus, it remains to show that for all $b\neq\beta$ there exists a model $\tilde{M}\in\mathcal{M}$ with $\mathbb{P}_M=\mathbb{P}_{\tilde{M}}$ such that the generalization loss is bounded below uniformly by a positive constant. We will show the stronger statement that for any $b\neq\beta$, there exists a model $\tilde{M}$ with $\mathbb{P}_{\tilde{M}}=\mathbb{P}_M$, such that under $\tilde{M}$, b results in arbitrarily large generalization error. Let $c>0$ and $i_0\in\mathcal{I}$. Define

$$\tilde{\sigma}:=\frac{\operatorname{sign}\left((\beta-b)i_0\right)\sqrt{1+c}-1}{(\beta-b)i_0}>0,$$

and let $\tilde{M}:=M(\gamma,\beta,\tilde{\sigma},Q)$. By construction of the model class $\mathcal{M}$, it holds that $\mathbb{P}_{\tilde{M}}=\mathbb{P}_M$. Furthermore, by an explicit computation we get that

$$\begin{aligned}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\Big[(Y-bX)^2\Big] &\geq \mathbb{E}_{\tilde{M}(i_0)}\Big[(Y-bX)^2\Big]\\ &=\mathbb{E}_{\tilde{M}(i_0)}\Big[((\beta-b)i_0H+\varepsilon_Y+\tfrac{1}{\tilde{\sigma}}H)^2\Big]\\ &=\mathbb{E}_{\tilde{M}(i_0)}\Big[([(\beta-b)i_0\tilde{\sigma}+1]\varepsilon_H+\varepsilon_Y)^2\Big]\\ &=[(\beta-b)i_0\tilde{\sigma}+1]^2+1\\ &=((\beta-b)i_0\tilde{\sigma})^2+2(\beta-b)i_0\tilde{\sigma}+2\\ &=(\operatorname{sign}\left((\beta-b)i_0\right)\sqrt{1+c}-1)^2\\ &\quad+2\operatorname{sign}\left((\beta-b)i_0\right)\sqrt{1+c}\\ &=c+2.\end{aligned}\tag{B.10}$$

Finally, by definition of the infimum, it holds that

$$\inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \beta X)^2\big] = 2. \tag{B.11}$$

Combining (B.10) and (B.11) yields that the generalization error is bounded below by c. That is,

$$\Big| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \Big| \geq c.$$

The above results make no assumptions on γ, and hold true, in particular, if $\gamma \neq 0$ (in which case Assumption 3.1 is satisfied, see Appendix B.2). This completes the proof of Proposition 3.6.

□

Proof of Proposition 3.7: Let $\mathcal{I}$ be a well-behaved set of interventions on X. We consider two cases; (A) all interventions in $\mathcal{I}$ are confounding-preserving and (B) there is at least one intervention in $\mathcal{I}$ that is confounding-removing.

Case (A): In this case, we prove the result in two steps: (i) We show that (A, ξ_X, ξ_Y) is identified from the observational distribution $\mathbb{P}_M$. (ii) We show that this implies that the intervention distributions (X^i, Y^i), $i \in \mathcal{I}$, are also identified from the observational distribution, and conclude by using Proposition 3.5. Some of the details will be slightly technical because we allow for a large class of distributions (e.g., there is no assumption on the existence of densities).

We begin with step (i). In this case, $\mathcal{I}$ is a set of confounding-preserving interventions on X, and we have that $\text{supp}_{\mathcal{I}}(X) \subseteq \text{supp}(X)$. Fix $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$ and let $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ be generated by the SCM of $\tilde{M}$. We have that $(X, Y, A) \overset{\mathcal{D}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ and by Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\text{supp}(X)$, hence $f(X) \overset{\text{a.s}}{=} \tilde{f}(X)$. Further, fix any $B \in \mathcal{B}(\mathbb{R}^p)$ (i.e., in the Borel sigma-algebra on $\mathbb{R}^p$) and note that

$$\begin{aligned}\mathbb{E}_M[\mathbb{1}_B(A)X|A] &= \mathbb{E}_M[\mathbb{1}_B(A)g(A) + \mathbb{1}_B(A)h_2(H, \varepsilon_X)|A] \\ &= \mathbb{E}_M[\mathbb{1}_B(A)g(A)|A] + \mathbb{1}_B(A)\mathbb{E}[h_2(H, \varepsilon_X)] = \mathbb{1}_B(A)g(A),\end{aligned}$$

almost surely. Here, we have used our modeling assumption $\mathbb{E}[h_2(H, \varepsilon_X)] = 0$. Hence, by similar arguments for $\mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A})$ and the fact that $(X, Y, A) \overset{\mathcal{D}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ we have that

$$\begin{aligned}\mathbb{1}_B(A)g(A) &\overset{\text{a.s}}{=} \mathbb{E}_M(\mathbb{1}_B(A)X|A) \\ &\overset{\mathcal{D}}{=} \mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A}) \\ &\overset{\text{a.s}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A}).\end{aligned}$$

We conclude that $\mathbb{1}_B(A)g(A) \overset{\mathcal{D}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A})$ for any $B \in \mathcal{B}(\mathbb{R}^p)$. Let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ denote the respective background probability measures on which the random

elements (X, Y, H, A) and $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ are defined. Fix any $F \in \sigma(A)$ (i.e., in the sigma-algebra generated by A) and note that there exists a $B \in \mathcal{B}(\mathbb{R}^p)$ such that $F = \{A \in B\}$. Since $A \overset{\mathcal{D}}{=} \tilde{A}$, we have that,

$$\begin{aligned}\int_F g(A)\,\mathrm{d}\mathbb{P} &= \int \mathbb{1}_B(A) g(A)\,\mathrm{d}\mathbb{P} \\ &= \int \mathbb{1}_B(\tilde{A}) \tilde{g}(\tilde{A})\,\mathrm{d}\tilde{\mathbb{P}} \\ &= \int \mathbb{1}_B(A) \tilde{g}(A)\,\mathrm{d}\mathbb{P} \\ &= \int_F \tilde{g}(A)\,\mathrm{d}\mathbb{P}.\end{aligned}$$

Both $g(A)$ and $\tilde{g}(A)$ are $\sigma(A)$-measurable and they agree integral-wise over every set $F \in \sigma(A)$, so we must have that $g(A) \overset{\text{a.s}}{=} \tilde{g}(A)$. With $\eta(a, b, c) = (a, c - \tilde{f}(b), b - \tilde{g}(a))$ we have that

$$\begin{aligned}(A, \xi_Y, \xi_X) &\overset{\text{a.s}}{=} (A, Y - \tilde{f}(X), X - \tilde{g}(A)) \\ &= \eta(A, X, Y) \\ &\overset{\mathcal{D}}{=} \eta(\tilde{A}, \tilde{X}, \tilde{Y}) \\ &= (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X),\end{aligned}$$

so $(A, \xi_Y, \xi_X) \overset{\mathcal{D}}{=} (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X)$. This completes step (i).

Next, we proceed with step (ii). Take an arbitrary intervention $i \in \mathcal{I}$ and let $\varphi^i, I^i, \tilde{I}^i$ with $I^i \overset{\mathcal{D}}{=} \tilde{I}^i$, $I^i \perp\!\!\!\perp (\varepsilon^i_X, \varepsilon^i_Y, \varepsilon^i_H, \varepsilon^i_A) \sim Q$ and $\tilde{I}^i \perp\!\!\!\perp (\tilde{\varepsilon}^i_X, \tilde{\varepsilon}^i_Y, \tilde{\varepsilon}^i_H, \tilde{\varepsilon}^i_A) \sim \tilde{Q}$ be such that the structural assignments for X^i and $\tilde{X}^i$ in $M(i)$ and $\tilde{M}(i)$, respectively, are given as

$$\begin{aligned}X^i &:= \varphi^i(A^i, g(A^i), h_2(H^i, \varepsilon^i_X), I^i), \\ \tilde{X}^i &:= \varphi^i(\tilde{A}^i, \tilde{g}(\tilde{A}^i), \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}^i_X), \tilde{I}^i).\end{aligned}$$

Define $\xi^i_X := h_2(H^i, \varepsilon^i_X)$, $\xi^i_Y := h_1(H^i, \varepsilon^i_Y)$, $\tilde{\xi}^i_X := \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}^i_X)$ and $\tilde{\xi}^i_Y := \tilde{h}_1(\tilde{H}^i, \tilde{\varepsilon}^i_Y)$. Then, it holds that

$$(A^i, \xi^i_X, \xi^i_Y) \overset{\mathcal{D}}{=} (A, \xi_X, \xi_Y) \overset{\mathcal{D}}{=} (\tilde{A}, \tilde{\xi}_X, \tilde{\xi}_Y) \overset{\mathcal{D}}{=} (\tilde{A}^i, \tilde{\xi}^i_X, \tilde{\xi}^i_Y),$$

where we used step (i), that (A^i, ξ^i_X, ξ^i_Y) and (A, ξ_X, ξ_Y) are generated by identical functions of the noise innovations and that $(\varepsilon_X, \varepsilon_Y, \varepsilon_H, \varepsilon_A)$ and $(\varepsilon^i_X, \varepsilon^i_Y, \varepsilon^i_H, \varepsilon^i_A)$ have identical distributions. Adding a random variable with the same distribution, that is mutually independent with all other variables, on both sides does not change the distribution of the bundle, hence

$$(A^i, \xi^i_X, \xi^i_Y, I^i) \overset{\mathcal{D}}{=} (\tilde{A}^i, \tilde{\xi}^i_X, \tilde{\xi}^i_Y, \tilde{I}^i).$$

Define $\kappa(a, b, c, d) := (\varphi^i(a, \tilde{g}(a), b, d), \tilde{f}(\varphi^i(a, \tilde{g}(a), b, d)) + c)$. As shown in step (i) above, we have that $g(A^i) \overset{\text{a.s}}{=} \tilde{g}(A^i)$. Furthermore, since $\operatorname{supp}(X^i) \subseteq \operatorname{supp}(X)$ we

have that $f(X^i) \overset{\text{a.s}}{=} \tilde{f}(X^i)$, and hence

$$\begin{aligned}(X^i, Y^i) &\overset{\text{a.s}}{=} (X^i, \tilde{f}(X^i) + \xi_Y^i) \\ &= (\varphi^i(A^i, g(A^i), \xi_X^i, I^i),\ \tilde{f}(\varphi^i(A^i, g(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\ &\overset{\text{a.s}}{=} (\varphi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i),\ \tilde{f}(\varphi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\ &= \kappa(A^i, \xi_X^i, \xi_Y^i, I^i) \\ &\overset{\mathcal{D}}{=} \kappa(\tilde{A}^i, \tilde{\xi}_X^i, \tilde{\xi}_Y^i, \tilde{I}^i) \\ &= (\tilde{X}^i, \tilde{Y}^i).\end{aligned}$$

Thus, $\mathbb{P}^{(X,Y)}_{M(i)} = \mathbb{P}^{(X,Y)}_{\tilde{M}(i)}$, which completes step (ii). Since $i \in \mathcal{I}$ was arbitrary, the result now follows from Proposition 3.5.

Case (B): Assume that the intervention set $\mathcal{I}$ contains at least one confounding-removing intervention. Let $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Then, by Proposition 3.1, it follows that the causal function $\tilde{f}$ is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. By Assumption 3.1, we further have that $\tilde{f}$ and f coincide on $\text{supp}(X) \supseteq \text{supp}_{\mathcal{I}}(X)$. Hence, it follows that

$$\begin{aligned}\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - \tilde{f}(X))^2] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f(X))^2],\end{aligned}$$

showing that also f is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. This completes the proof of Proposition 3.7. □

Proof of Proposition 3.8: Let $\tilde{M} \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By Assumptions 3.1 and 3.2, it holds that $f \equiv \tilde{f}$. The proof now proceeds analogously to that of Proposition 3.7. □

Proof of Proposition 3.9: By Assumption 3.1, f is identified on $\text{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of interventions containing at least one confounding-removing intervention. For any $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$, Proposition 3.1 yields that the causal function is a minimax solution. That is,

$$\begin{aligned}\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \tilde{f}(X))^2\right] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] = \mathbb{E}_{\tilde{M}}[\xi_Y^2], \qquad \text{(B.12)}\end{aligned}$$

where we used that any intervention $i \in \mathcal{I}$ does not affect the distribution of $\xi_Y = \tilde{h}_2(H, \varepsilon_Y)$. Now, assume that $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ satisfies $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\text{supp}^M(X) = \text{supp}^{\tilde{M}}(X)$. Let f^* be any function in $\mathcal{F}$ such that $f^* = f$ on $\text{supp}^M(X)$. We first show that $\|\tilde{f} - f^*\|_{\mathcal{I},\infty} \leq 2\delta K$, where $\|f\|_{\mathcal{I},\infty} := \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \|f(x)\|$. By the mean value

theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x-y\|$, for all $x, y \in \mathcal{D}$. For any $x \in \text{supp}^M_{\mathcal{I}}(X)$ and $y \in \text{supp}^M(X)$ we have

$$\begin{aligned}\left|\tilde{f}(x) - f^*(x)\right| &= \left|\tilde{f}(x) - \tilde{f}(y) + f^*(y) - f^*(x)\right| \\ &\leq \left|\tilde{f}(x) - \tilde{f}(y)\right| + \left|f^*(y) - f^*(x)\right| \\ &\leq 2K\|x-y\|,\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y) = f^*(y)$, for all $y \in \text{supp}^M(X)$. In particular, it holds that

$$\begin{aligned}\|\tilde{f} - f^*\|_{\mathcal{I},\infty} &= \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \left|\tilde{f}(x) - f^*(x)\right| \\ &\leq 2K \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \inf_{y \in \text{supp}^M(X)} \|x-y\| \\ &= 2\delta K.\end{aligned} \tag{B.13}$$

For any $i \in \mathcal{I}$ we have that

$$\begin{aligned}\mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] &= \mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) + \xi_Y - f^*(X))^2\right] \\ &= \mathbb{E}_{\tilde{M}}\left[\xi_Y^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f^*(X))^2\right] \\ &\quad + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f^*(X))\right].\end{aligned} \tag{B.14}$$

Next, we can use Cauchy-Schwarz, (B.12) and (B.13) in (B.14) to get that

$$\begin{aligned}&\left|\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\right| \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \mathbb{E}_{\tilde{M}}[\xi_Y^2] \\ &= \sup_{i \in \mathcal{I}} \left(\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f^*(X))^2\right] + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f^*(X))\right]\right) \\ &\leq 4\delta^2 K^2 + 4\delta K\sqrt{\text{Var}_M(\xi_Y)},\end{aligned} \tag{B.15}$$

proving the first statement. Finally, if $\mathcal{I}$ consists only of confounding-removing interventions, then the bound in (B.15) can be improved by using that $\mathbb{E}[\xi_Y] = 0$ together with $H \perp\!\!\!\perp X$. In that case, we get that $\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f(X))\right] = 0$ and hence the bound becomes $4\delta^2 K^2$. This completes the proof of Proposition 3.9. □

Proof of Proposition 3.10: By Assumption 3.1, f is identified on $\text{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of confounding-preserving interventions. For a fixed $\varepsilon > 0$, let $f^* \in \mathcal{F}$ be a function satisfying

$$\left|\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f^*(X))^2\right) - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right)\right| \leq \varepsilon. \tag{B.16}$$

Fix any secondary model $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. The general idea is to derive an upper bound for $\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f^*(X))^2]$ and a lower bound

for $\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2]$ which will allow us to bound the absolute difference of interest.

Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\text{supp}^M(X) = \text{supp}^{\tilde{M}}(X)$. We first show that

$$\|\tilde{f} - f\|_{\mathcal{I},\infty} \leq 2\delta K,$$

where $\|f\|_{\mathcal{I},\infty} := \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \|f(x)\|$. By the mean value theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x - y\|$, for all $x, y \in \mathcal{D}$. For any $x \in \text{supp}_{\mathcal{I}}^M(X)$ and $y \in \text{supp}^M(X)$ we have

$$\begin{aligned}\left|\tilde{f}(x) - f(x)\right| &= \left|\tilde{f}(x) - \tilde{f}(y) + f(y) - f(x)\right| \\ &\leq \left|\tilde{f}(x) - \tilde{f}(y)\right| + \left|f(y) - f(x)\right| \\ &\leq 2K\|x - y\|,\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y)$, for all $y \in \text{supp}_M(X)$. In particular, it holds that

$$\begin{aligned}\|\tilde{f} - f\|_{\mathcal{I},\infty} &= \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \left|\tilde{f}(x) - f(x)\right| \\ &\leq 2K \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \inf_{y \in \text{supp}^M(X)} \|x - y\| \\ &= 2\delta K.\end{aligned} \tag{B.17}$$

Let now $i \in \mathcal{I}$ be fixed. The term $\xi_Y = h_1(H, \varepsilon_Y)$ is not affected by the intervention i. Furthermore, $\mathbb{P}_{M(i)}^{(X,\xi_Y)} = \mathbb{P}_{\tilde{M}(i)}^{(X,\xi_Y)}$ since i is confounding-preserving (this can be seen by a slight modification to the arguments from case (A) in the proof of Proposition 3.7). Thus, for any $f_\diamond \in \mathcal{F}$ we have that

$$\begin{aligned}&\mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \\ &= \mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) + \xi_Y - f_\diamond(X) + f(X) - f(X))^2\right] \\ &= \mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[(f(X) - f_\diamond(X))^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))^2\right] \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(f(X) - f_\diamond(X))\right] \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\right] \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f(X))\right] \\ &= \mathbb{E}_{M(i)}\left[\xi_Y^2\right] + \mathbb{E}_{M(i)}\left[(f(X) - f_\diamond(X))^2\right] + \mathbb{E}_{M(i)}\left[(\tilde{f}(X) - f(X))^2\right] \\ &\qquad + 2\mathbb{E}_{M(i)}\left[\xi_Y(f(X) - f_\diamond(X))\right] \\ &\qquad + 2\mathbb{E}_{M(i)}\left[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\right] \\ &\qquad + 2\mathbb{E}_{M(i)}\left[\xi_Y(\tilde{f}(X) - f(X))\right] \\ &= \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f_\diamond) + L_3^i(\tilde{f}),\end{aligned} \tag{B.18}$$

where, we have made the following definitions

$$\begin{aligned} L_1^i(\tilde{f}) &\coloneqq \mathbb{E}_{M(i)}\Big[(\tilde{f}(X) - f(X))^2\Big], \\ L_2^i(\tilde{f}, f_\diamond) &\coloneqq 2\mathbb{E}_{M(i)}\Big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\Big], \\ L_3^i(\tilde{f}) &\coloneqq 2\mathbb{E}_{M(i)}\Big[\xi_Y(\tilde{f}(X) - f(X))\Big]. \end{aligned}$$

Using (B.17) it follows that

$$0 \leq L_1^i(\tilde{f}) \leq 4\delta^2 K^2, \tag{B.19}$$

and by the Cauchy-Schwarz inequality it follows that

$$\Big|L_3^i(\tilde{f})\Big| \leq 2\sqrt{\mathrm{Var}_M(\xi_Y)4\delta^2K^2} = 4\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}. \tag{B.20}$$

Let now $f_\diamond \in \mathcal{F}$ be any function such that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_\diamond(X))^2)\Big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - \tilde{f}(X))^2)\Big], \tag{B.21}$$

then by (B.17), the Cauchy-Schwarz inequality and Proposition 3.3, it holds for all $i \in \mathcal{I}$ that

$$\begin{aligned} L_2^i(\tilde{f}, f_\diamond) &= 2\mathbb{E}_{M(i)}\Big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\Big] \\ &= 2\mathbb{E}_{\tilde{M}(i)}\Big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\Big] \\ &= -2\mathbb{E}_{\tilde{M}(i)}\Big[(\tilde{f}(X) - f(X))^2\Big] \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}\Big[(\tilde{f}(X) - f(X))(\tilde{f}(X) - f_\diamond(X))\Big] \qquad \text{(B.22)} \\ &\geq -8\delta^2K^2 - 2\sqrt{4\delta^2K^2}\sqrt{4\,\mathrm{Var}_M(\xi_Y)} \\ &= -8\delta^2K^2 - 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}, \qquad \text{(B.23)} \end{aligned}$$

where, in the third equality, we have added and subtracted the term $2\mathbb{E}_{\tilde{M}(i)}\Big[(\tilde{f}(X) - f(X))\tilde{f}(X)\Big]$. Now let

$$\mathcal{S} \coloneqq \{f_\diamond \in \mathcal{F} : \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_\diamond(X))^2\Big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - \tilde{f}(X))^2\Big]\}$$

be the set of all functions satisfying (B.21). Due to (B.18), (B.19), (B.20) and (B.23) we have the following lower bound of interest

$$\begin{aligned} &\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_\diamond(X))^2\Big] \\ &= \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - f_\diamond(X))^2\Big] \\ &= \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \Big\{\mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f_\diamond) + L_3^i(\tilde{f})\Big\} \\ &\geq \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big] - 8\delta^2K^2 - 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\ &\qquad - 4\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \qquad \text{(B.24)} \\ &\geq \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big] - 8\delta^2K^2 - 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}. \qquad \text{(B.25)} \end{aligned}$$

Next, we construct the aforementioned upper bound of interest. To that end, note that

$$\begin{aligned}&\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\Big[(Y-f^*(X))^2\Big]\\&\quad=\sup_{i\in\mathcal{I}}\Big\{\mathbb{E}_{M(i)}\Big[(Y-f^*(X))^2\Big]+L_1^i(\tilde{f})+L_2^i(\tilde{f},f^*)+L_3^i(\tilde{f})\Big\},\end{aligned}\tag{B.26}$$

by (B.18). We have already established upper bounds for $L_1^i(\tilde{f})$ and $L_3^i(\tilde{f})$ in (B.19) and (B.20), respectively. In order to control $L_2^i(\tilde{f},f^*)$ we introduce an auxiliary function. Let $\bar{f}^*\in\mathcal{F}$ satisfy

$$\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-\bar{f}^*(X))^2)\Big]\leq\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-f(X))^2)\Big],\tag{B.27}$$

and

$$\left|\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-\bar{f}^*(X))^2\Big]-\inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-f_\diamond(X))^2\Big]\right|\leq\varepsilon.\tag{B.28}$$

Choosing such a $\bar{f}^*\in\mathcal{F}$ is always possible. If f is an ε-minimax solution, i.e., it satisfies (B.28), then choose $\bar{f}^*=f$. Otherwise, if f is not a ε-minimax solution, then choose any $\bar{f}^*\in\mathcal{F}$ that is an ε-minimax solution (which is always possible). In this case we have that

$$\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-\bar{f}^*(X))^2\Big]-\inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-f_\diamond(X))^2\Big]\leq\varepsilon,$$

and

$$\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-f(X))^2\Big]-\inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}}\mathbb{E}_{M(i)}\Big[(Y-f_\diamond(X))^2\Big]\geq\varepsilon,$$

which implies that (B.27) is satisfied. We can now construct an upper bound on $L_2^i(\tilde{f},f^*)$ in terms of $L_2^i(\tilde{f},\bar{f}^*)$ by noting that for all $i\in\mathcal{I}$

$$\begin{aligned}\Big|L_2^i(\tilde{f},f^*)\Big|&=2\Big|\mathbb{E}_{M(i)}\Big[(\tilde{f}(X)-f(X))(f(X)-f^*(X))\Big]\Big|\\&\leq 2\Big|\mathbb{E}_{M(i)}\Big[(\tilde{f}(X)-f(X))(f(X)-\bar{f}^*(X))\Big]\Big|\\&\quad+2\mathbb{E}_{M(i)}\Big|(\tilde{f}(X)-f(X))(\bar{f}^*(X)-f^*(X))\Big|\\&=\Big|L_2^i(\tilde{f},\bar{f}^*)\Big|+2\mathbb{E}_{M(i)}\Big|(\tilde{f}(X)-f(X))(\bar{f}^*(X)-f^*(X))\Big|\\&\leq 2\sqrt{\mathbb{E}_{M(i)}\left[(\tilde{f}(X)-f(X))^2\right]\mathbb{E}_{M(i)}\left[(\bar{f}^*(X)-f^*(X))^2\right]}\\&\quad+\Big|L_2^i(\tilde{f},\bar{f}^*)\Big|\qquad\text{(B.29)}\\&\leq\Big|L_2^i(\tilde{f},\bar{f}^*)\Big|+4\delta K\sqrt{\mathbb{E}_{M(i)}\Big[(\bar{f}^*(X)-f^*(X))^2\Big]},\qquad\text{(B.30)}\end{aligned}$$

where we used the triangle inequality, Cauchy-Schwarz inequality and (B.17). Furthermore, (B.17) and (B.27) together with Proposition 3.3 yield the following bound

$$\begin{aligned}|L_2^i(\tilde{f}, \bar{f}^*)| &= 2\Big|\mathbb{E}_{M(i)}\Big[(\tilde{f}(X) - f(X))(f(X) - \bar{f}^*(X))\Big]\Big| \\ &= 2\sqrt{\mathbb{E}_{M(i)}\Big[(\tilde{f}(X) - f(X))^2\Big]\mathbb{E}_{M(i)}\Big[(f(X) - \bar{f}^*(X))^2\Big]} \\ &\leq 2\sqrt{4\delta^2 K^2}\sqrt{4\operatorname{Var}_M(\xi_Y)} \\ &= 8\delta K\sqrt{\operatorname{Var}_M(\xi_Y)}, \end{aligned} \tag{B.31}$$

for any $i \in \mathcal{I}$. Thus, it suffices to construct an upper bound on the second term in the final expression in (B.30). Direct computation leads to

$$\begin{aligned}\mathbb{E}_{M(i)}\Big[(Y - f^*(X))^2\Big] = \mathbb{E}_{M(i)}&\Big[(Y - \bar{f}^*(X))^2\Big] \\ &+ \mathbb{E}_{M(i)}\Big[(\bar{f}^*(X) - f^*(X))^2\Big] \\ &+ 2\mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\Big].\end{aligned}$$

Rearranging the terms and applying the triangle inequality and Cauchy-Schwarz results in

$$\begin{aligned}&\mathbb{E}_{M(i)}\left[(\bar{f}^*(X) - f^*(X))^2\right] \\ &\quad= \mathbb{E}_{M(i)}\Big[(Y - f^*(X))^2\Big] - \mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))^2\Big] \\ &\qquad - 2\mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\Big] \\ &\quad\leq \Big|\mathbb{E}_{M(i)}\Big[(Y - f^*(X))^2\Big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big]\Big| \\ &\qquad + \Big|\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f_\diamond(X))^2\Big] - \mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))^2\Big]\Big| \\ &\qquad + 2\mathbb{E}_{M(i)}\Big|(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\Big| \\ &\quad\leq 2\varepsilon + 2\sqrt{\mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))^2\Big]}\sqrt{\mathbb{E}_{M(i)}\Big[(\bar{f}^*(X) - f^*(X))^2\Big]} \\ &\quad\leq 2\varepsilon + 2\sqrt{\operatorname{Var}_M(\xi_Y)}\sqrt{\mathbb{E}_{M(i)}\Big[(\bar{f}^*(X) - f^*(X))^2\Big]},\end{aligned}$$

for any $i \in \mathcal{I}$. Here, we used that both f^* and $\bar{f}^*$ are ε-minimax solutions with respect to M and that $\bar{f}^*$ satisfies (B.27) which implies that

$$\begin{aligned}\mathbb{E}_{M(i)}\Big[(Y - \bar{f}^*(X))^2)\Big] &\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[(Y - f(X))^2)\Big] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\Big[\xi_Y^2\Big] \\ &= \operatorname{Var}_M(\xi_Y),\end{aligned}$$

for any $i \in \mathcal{I}$, as ξ_Y is unaffected by an intervention on X. Thus, $\mathbb{E}_{M(i)}\Big[(\bar{f}^*(X) - f^*(X))^2\Big]$ must satisfy $\ell(\mathbb{E}_{M(i)}\Big[(\bar{f}^*(X) - f^*(X))^2\Big]) \leq 0$, where $\ell : [0, \infty) \to \mathbb{R}$ is

given by $\ell(z) = z - 2\varepsilon - 2\sqrt{\mathrm{Var}_M(\xi_Y)}\sqrt{z}$. The linear term of ℓ grows faster than the square root term, so the largest allowed value of $\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]$ coincides with the largest root of $\ell(z)$. The largest root is given by

$$C^2 := 2\varepsilon + 2\,\mathrm{Var}_M(\xi_Y) + 2\sqrt{\mathrm{Var}_M(\xi_Y)^2 + 2\varepsilon\,\mathrm{Var}_M(\xi_Y)},$$

where $(\cdot)^2$ refers to the square of C. Hence, for any $i \in \mathcal{I}$ it holds that

$$\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \leq C^2. \tag{B.32}$$

Hence by (B.30), (B.31) and (B.32) we have that the following upper bound is valid for any $i \in \mathcal{I}$.

$$\big|L_2^i(\tilde{f}, f^*)\big| \leq 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta KC. \tag{B.33}$$

Thus, using (B.26) with (B.19), (B.20) and (B.33), we get the following upper bound

$$\begin{aligned}
&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] \\
&\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] + 4\delta^2K^2 + 4\delta KC + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}.
\end{aligned} \tag{B.34}$$

Finally, by combining the bounds (B.25) and (B.34) together with (B.16) we get that

$$\begin{aligned}
&\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\right| \\
&\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \\
&\qquad + 4\delta^2K^2 + 4\delta KC + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\qquad + 8\delta^2K^2 + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\quad\leq \varepsilon + 12\delta^2K^2 + 24\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta KC.
\end{aligned} \tag{B.35}$$

Using that all terms are positive, we get that

$$C = \sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{\mathrm{Var}_M(\xi_Y) + 2\varepsilon} \leq 2\sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{2\varepsilon}$$

Hence, (B.35) is bounded above by

$$\varepsilon + 12\delta^2K^2 + 32\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K\sqrt{\varepsilon}.$$

This completes the proof of Proposition 3.10. □

Proof of Proposition 3.11: Let $\bar{f} \in \mathcal{F}$ and $c > 0$. By assumption, $\mathcal{I}$ is a well-behaved set of support-extending interventions on X. Since $\mathrm{supp}^M_{\mathcal{I}}(X) \setminus \mathrm{supp}^M(X)$

has non-empty interior, there exists an intervention $i_0 \in \mathcal{I}$ and $\varepsilon > 0$ such that $\mathbb{P}_{M(i_0)}(X \in B) \geq \varepsilon$, for some open subset $B \subsetneq \bar{B}$, such that $\operatorname{dist}(B, \mathbb{R}^d \setminus \bar{B}) > 0$, where $\bar{B} := \operatorname{supp}^M_{\mathcal{I}}(X) \setminus \operatorname{supp}^M(X)$. Let $\tilde{f}$ be any continuous function satisfying that, for all $x \in B \cup (\mathbb{R}^d \setminus \bar{B})$,

$$\tilde{f}(x) = \begin{cases} \bar{f}(x) + \gamma, & x \in B \\ f(x), & x \in \mathbb{R}^d \setminus \bar{B}, \end{cases}$$

where $\gamma := \varepsilon^{-1/2} \left\{ (2\mathbb{E}_{\tilde{M}}[\xi_Y^2] + c)^{1/2} + (\mathbb{E}_{\tilde{M}}[\xi_Y^2])^{1/2} \right\}$.

Consider a secondary model $\tilde{M} = (\tilde{f}, g, h_1, h_2, Q) \in \mathcal{M}$. Then, by Assumption 3.1, it holds that $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. Since $\mathcal{I}$ only consists of interventions on X, it holds that $\mathbb{P}_{M(i_0)}(X \in B) = \mathbb{P}_{\tilde{M}(i_0)}(X \in B)$ (this holds since all components of $\tilde{M}$ and M are equal, except for the function f, which is not allowed to enter in the intervention on X). Therefore,

$$\begin{aligned} \mathbb{E}_{\tilde{M}(i_0)}\left[(Y - \bar{f}(X))^2\right] &\geq \mathbb{E}_{\tilde{M}(i_0)}\left[(Y - \bar{f}(X))^2 \mathbb{1}_B(X)\right] \\ &= \mathbb{E}_{\tilde{M}(i_0)}\left[(\gamma + \xi_Y)^2 \mathbb{1}_B(X)\right] \\ &\geq \gamma^2 \varepsilon + 2\gamma \mathbb{E}_{\tilde{M}(i_0)}\left[\xi_Y \mathbb{1}_B(X)\right] \\ &\geq \gamma^2 \varepsilon - 2\gamma \left(\mathbb{E}_{\tilde{M}}\left[\xi_Y^2\right] \varepsilon\right)^{1/2} \\ &= c + \mathbb{E}_{\tilde{M}}[\xi_Y^2], \end{aligned} \tag{B.36}$$

where the third inequality follows from Cauchy–Schwarz. Further, by the definition of the infimum it holds that

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \tilde{f}(X))^2\right] = \mathbb{E}_{\tilde{M}}[\xi_Y^2]. \tag{B.37}$$

Therefore, combining (B.36) and (B.37), the claim follows. □

Proof of Proposition 3.12: We prove the result by showing that under Assumption 3.3 it is possible to express interventions on A as confounding-preserving interventions on X and applying Propositions 3.7 and 3.8. To avoid confusion, we will throughout this proof denote the true model by $M^0 = (f^0, g^0, h_1^0, h_2^0, Q^0)$. Fix an intervention $i \in \mathcal{I}$. Since it is an intervention on A, there exist ψ^i and I^i such that for any $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, the intervened SCM $M(i)$ is of the form

$$\begin{aligned} A^i &:= \psi^i(I^i, \varepsilon_A^i), \quad H^i := \varepsilon_H^i, \\ X^i &:= g(A^i) + h_2(H^i, \varepsilon_X^i), \\ Y^i &:= f(X^i) + h_1(H^i, \varepsilon_Y^i), \end{aligned}$$

where $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$. We now define a confounding-preserving intervention j on X, such that, for all models $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, the distribution of (X, Y) under $\tilde{M}(j)$ coincides with that under $\tilde{M}(i)$. To that end, define the intervention function

$$\bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j) := g^0(\psi^i(I^j, A^j)) + h_2(H^j, \varepsilon_X^j),$$

where g^0 is the fixed function corresponding to model M, and therefore not an argument of $\bar{\psi}^j$. Let now j be the intervention on X satisfying that, for all $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, the intervened model $M(j)$ is given as

$$\begin{aligned} A^j &:= \varepsilon_A^j, \quad H^j := \varepsilon_H^j, \\ X^j &:= \bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j), \\ Y^j &:= f(X^j) + h_1(H^j, \varepsilon_Y^j), \end{aligned}$$

where $(\varepsilon_X^j, \varepsilon_Y^j, \varepsilon_A^j, \varepsilon_H^j) \sim Q$ and where I^j is chosen such that $I^j \overset{\mathcal{D}}{=} I^i$. By definition, j is a confounding-preserving intervention. Let now $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q})$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and let $(\tilde{X}^i, \tilde{Y}^i)$ and $(\tilde{X}^j, \tilde{Y}^j)$ be generated under $\tilde{M}(i)$ and $\tilde{M}(j)$, respectively. By Assumption 3.3, it holds for all $a \in \text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$ that $\tilde{g}(a) = g^0(a)$. Hence, we get that

$$\begin{aligned} (\tilde{X}^i, \tilde{Y}^i) \overset{\mathcal{D}}{=} &(\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \tilde{f}(\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) \\ &+ \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\ = &(g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \tilde{f}(g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) \\ &+ \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\ \overset{\mathcal{D}}{=} &(g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) + \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j), \tilde{f}(g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) \\ &+ \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j)) + \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\ \overset{\mathcal{D}}{=} &(\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j), \tilde{f}(\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j)) \\ &+ \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\ \overset{\mathcal{D}}{=} &(\tilde{X}^j, \tilde{Y}^j), \end{aligned}$$

as desired. Since $i \in \mathcal{I}$ was arbitrary, we have now shown that there exists a mapping π from $\mathcal{I}$ into a set $\mathcal{J}$ of confounding-preserving (and hence a well-behaved set) of interventions on X, such that for all $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, $\mathbb{P}^{(X,Y)}_{\tilde{M}(i)} = \mathbb{P}^{(X,Y)}_{\tilde{M}(\pi(i))}$. Hence, we can rewrite Equation (3.3) in Definition 3.1 in terms of the set $\mathcal{J}$. The result now follows from Propositions 3.7 and 3.8. □

Proof of Proposition 3.13: Let $b \in \mathbb{R}^d$ be such that $f(x) = b^\top x$ for all $x \in \mathbb{R}^d$. We start by characterizing the error $\mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]$. Let us consider models of the form $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$ for some function $\tilde{g} \in \mathcal{G}$ with $\tilde{g}(a) = g(a)$ for all $a \in \text{supp}_M(A)$. Clearly, any such model satisfies that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. For every $a \in \mathcal{A}$, let $i_a \in \mathcal{I}$ denote the corresponding hard intervention on A. For every

$a \in \mathcal{A}$ and $b_\diamond \in \mathbb{R}^d$, we then have

$$\begin{aligned}
&\mathbb{E}_{\tilde{M}(i_a)}\Big[(Y - b_\diamond^\top X)^2\Big] \\
&= \mathbb{E}_{\tilde{M}(i_a)}\Big[(b^\top X + \xi_Y - b_\diamond^\top X)^2\Big] \\
&= (b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[XX^\top](b - b_\diamond) \\
&\quad + 2(b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[X\xi_Y] + \mathbb{E}_{\tilde{M}(i_a)}\Big[\xi_Y^2\Big] \\
&= (b - b_\diamond)^\top \underbrace{(\tilde{g}(a)\tilde{g}(a)^\top + \mathbb{E}_M[\xi_X \xi_X^\top])}_{=:K_{\tilde{M}}(a)}(b - b_\diamond) \\
&\quad + 2(b - b_\diamond)^\top \mathbb{E}_M[\xi_X \xi_Y] + \mathbb{E}_M\Big[\xi_Y^2\Big],
\end{aligned} \tag{B.38}$$

where we have used that, under i_a, the distribution of (ξ_X, ξ_Y) is unaffected. We now show that, for any $\tilde{M}$ with the above form, the causal function f does not minimize the worst-case risk across interventions in $\mathcal{I}$. The idea is to show that the worst-case risk (B.38) strictly decreases at $b_\diamond = b$ in the direction $u := \mathbb{E}_M[\xi_X \xi_Y]/\|\mathbb{E}_M[\xi_X \xi_Y]\|_2$. For every $a \in \mathcal{A}$ and $s \in \mathbb{R}$, define

$$\begin{aligned}
\ell_{\tilde{M},a}(s) :&= \mathbb{E}_{\tilde{M}(i_a)}\Big[(Y - (b + su)^\top X)^2\Big] \\
&= u^\top K_{\tilde{M}}(a) u \cdot s^2 - 2u^\top \mathbb{E}_M[\xi_X \xi_Y] \cdot s + \mathbb{E}_M\Big[\xi_Y^2\Big].
\end{aligned}$$

For every a, $\ell'_{\tilde{M},a}(0) = -2\|\mathbb{E}_M[\xi_X \xi_Y]\|_2 < 0$, showing that $\ell_{\tilde{M},a}$ is strictly decreasing at $s = 0$ (with a derivative that is bounded away from 0 across all $a \in \mathcal{A}$). By boundedness of $\mathcal{A}$ and by the continuity of $a \mapsto \ell''_{\tilde{M},a}(0) = 2u^\top K_{\tilde{M}}(a) u$, it further follows that $\sup_{a \in \mathcal{A}} |\ell''_{\tilde{M},a}(0)| < \infty$. Hence, we can find $s_0 > 0$ such that for all $a \in \mathcal{A}$, $\ell_{\tilde{M},a}(0) > \ell_{\tilde{M},a}(s_0)$. It now follows by continuity of $(a, s) \mapsto \ell_{\tilde{M},a}(s)$ that

$$\begin{aligned}
\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - b^\top X)^2\Big] &= \sup_{a \in \mathcal{A}} \ell_{\tilde{M},a}(0) \\
&> \sup_{a \in \mathcal{A}} \ell_{\tilde{M},a}(s_0) \\
&= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\Big[(Y - (b + s_0 u)^\top X)^2\Big],
\end{aligned}$$

showing that $b + s_0 u$ attains a lower worst-case risk than b.

We now show that all functions other than f may result in an arbitrarily large error. Let $\bar{b} \in \mathbb{R}^d \setminus \{b\}$ be given, and let $j \in \{1, \ldots, d\}$ be such that $b_j \neq \bar{b}_j$. The idea is to construct a function $\tilde{g} \in \mathcal{G}$ such that, under the corresponding model $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$, some hard interventions on A result in strong shifts of the jth coordinate of X. Let $a \in \mathcal{A}$. Let $e_j \in \mathbb{R}^d$ denote the jth unit vector, and assume that $\tilde{g}(a) = ne_j$ for some $n \in \mathbb{N}$. Using (B.38), it follows that

$$\begin{aligned}
\mathbb{E}_{\tilde{M}(i_a)}\Big[(Y - \bar{b}^\top X)^2\Big] =& n^2(\bar{b}_j - b_j)^2 + (\bar{b} - b)^\top \mathbb{E}_M[\xi_X \xi_X^\top](\bar{b} - b) \\
&+ 2(\bar{b} - b)^\top \mathbb{E}_M[\xi_X \xi_Y] + \mathbb{E}_M\Big[\xi_Y^2\Big].
\end{aligned}$$

By letting $n \to \infty$, we see that the above error may become arbitrarily large. Given any $c > 0$, we can therefore construct $\tilde{g}$ such that $\mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] \geq c + \mathbb{E}_M\left[\xi_Y^2\right]$. By carefully choosing $a \in \operatorname{int}(\mathcal{A} \setminus \operatorname{supp}_M(A))$, this can be done such that $\tilde{g}$ is continuous and $\tilde{g}(a) = g(a)$ for all $a \in \operatorname{supp}_M(A)$, ensuring that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. It follows that

$$\begin{aligned}
c &\leq \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \mathbb{E}_M\left[\xi_Y^2\right] \\
&= \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b^\top X)^2\right] \\
&\leq \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \inf_{b_\diamond \in \mathbb{R}^d} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond^\top X)^2\right] \\
&\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \bar{b}^\top X)^2\right] - \inf_{b_\diamond \in \mathbb{R}^d} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond^\top X)^2\right],
\end{aligned}$$

which completes the proof of Proposition 3.13. □

Proof of Proposition 3.14: By assumption, $\mathcal{I}$ is a set of interventions on X or A of which at least one is confounding-removing. Now fix any

$$\tilde{M} = (f_{\eta_0}(x; \tilde{\theta}), \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M},$$

with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. By Proposition 3.1, we have that a minimax solution is given by the causal function. That is,

$$\begin{aligned}
\inf_{f_\diamond \in \mathcal{F}_{\eta_0}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X; \tilde{\theta}))^2\right] \\
&= \mathbb{E}_M[\xi_Y^2],
\end{aligned}$$

where we used that ξ_Y is unaffected by an intervention on X. By the support restriction $\operatorname{supp}^M(X) \subseteq (a, b)$ we know that

$$\begin{aligned}
f_{\eta_0}(x; \theta^0) &= B(x)^\top \theta^0, \\
f_{\eta_0}(x; \tilde{\theta}) &= B(x)^\top \tilde{\theta}, \\
f_{\eta_0}(x; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}) &= B(x)^\top \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu},
\end{aligned}$$

for all $x \in \operatorname{supp}^M(X)$. Furthermore, as $Y = B(X)^\top \theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that

$$\begin{aligned}
\mathbb{E}_M\left[C(A)Y\right] &= \mathbb{E}_M\left[C(A)B(X)^\top \theta^0\right] + \mathbb{E}_M\left[C(A)\xi_Y\right] \\
&= \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0, \qquad \text{(B.39)}
\end{aligned}$$

where we used the assumptions that $\mathbb{E}\left[\xi_Y\right] = 0$ and $A \perp\!\!\!\perp \xi_Y$ by the exogeneity of A. Similarly,

$$\mathbb{E}_{\tilde{M}}\left[C(A)Y\right] = \mathbb{E}_{\tilde{M}}\left[C(A)B(X)^\top\right]\tilde{\theta}.$$

As $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, we have that $\mathbb{E}_M[C(A)Y] = \mathbb{E}_{\tilde{M}}[C(A)Y]$ and $\mathbb{E}_M[C(A)B(X)^\top] = \mathbb{E}_{\tilde{M}}[C(A)B(X)^\top]$, hence

$$\mathbb{E}_M\left[C(A)B(X)^\top\right]\tilde{\theta} = \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0 \iff \tilde{\theta} = \theta^0,$$

by assumption (B2), which states that $\mathbb{E}[C(A)B(X)^\top]$ is of full rank (bijective). In other words, the causal function parameterized by θ^0 is identified from the observational distribution. Assumptions 3.1 and 3.2 are therefore satisfied. Furthermore, we also have that

$$\begin{aligned}
&\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] \\
&\quad= \sup_{i\in\mathcal{I}}\Big\{\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))\right]\Big\} \\
&\quad\leq \sup_{i\in\mathcal{I}}\Big\{\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right]\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right]}\Big\} \\
&\quad\leq \sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] + \mathbb{E}_M\left[\xi_Y^2\right] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_M\left[\xi_Y^2\right]\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right]},
\end{aligned}$$

by Cauchy-Schwarz inequality, where we additionally used that $\mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2]$ as ξ_Y is unaffected by interventions on X. Thus,

$$\begin{aligned}
&\Big|\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\Big| \\
&\quad\leq \sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] \\
&\qquad + 2\sqrt{\mathbb{E}_M\left[\xi_Y^2\right]\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right]}.
\end{aligned}$$

For the next few derivations let $\hat{\theta} = \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}$ for notational simplicity. Note that, for all $x \in \mathbb{R}$,

$$\begin{aligned}
(f_{\eta_0}(x;\theta^0) - f_{\eta_0}(x;\hat{\theta}))^2 \leq\ &(\theta^0 - \hat{\theta})^\top B(x)B(x)^\top(\theta^0 - \hat{\theta}) \\
&+ (B(a)^\top(\theta^0 - \hat{\theta}) + B'(a)^\top(\theta^0 - \hat{\theta})(x - a))^2 \\
&+ (B(b)^\top(\theta^0 - \hat{\theta}) + B'(b)^\top(\theta^0 - \hat{\theta})(x - b))^2.
\end{aligned}$$

The second term has the following upper bound

$$\begin{aligned}
&(B(a)^\top(\theta^0-\hat\theta)+B'(a)^\top(\theta^0-\hat\theta)(x-a))^2\\
&\quad=(\theta^0-\hat\theta)^\top B(a)B(a)^\top(\theta^0-\hat\theta)\\
&\qquad+(x-a)^2(\theta^0-\hat\theta)^\top B'(a)B'(a)^\top(\theta^0-\hat\theta)\\
&\qquad+2(x-a)(\theta^0-\hat\theta)^\top B'(a)B(a)^\top(\theta^0-\hat\theta)\\
&\quad\le\lambda_{\mathrm{m}}(B(a)B(a)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\qquad+(x-a)^2\lambda_{\mathrm{m}}(B'(a)B'(a)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\qquad+(x-a)\lambda_{\mathrm{m}}(B'(a)B(a)^\top+B(a)B'(a)^\top)\|\theta^0-\hat\theta\|_2^2,
\end{aligned}$$

where λ_{m} denotes the maximum eigenvalue. An analogous upper bound can be constructed for the third term. Thus, by combining these two upper bounds with a similar upper bound for the first term, we arrive at

$$\begin{aligned}
&\mathbb{E}_{\tilde M(i)}\big[(f_{\eta_0}(X;\theta^0)-f_{\eta_0}(X;\hat\theta))^2\big]\\
&\le\lambda_{\mathrm{m}}(\mathbb{E}_{\tilde M(i)}[B(X)B(X)^\top])\|\theta^0-\hat\theta\|_2^2\\
&\quad+\lambda_{\mathrm{m}}(B(a)B(a)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\quad+\mathbb{E}_{\tilde M(i)}[(X-a)^2]\lambda_{\mathrm{m}}(B'(a)B'(a)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\quad+\mathbb{E}_{\tilde M(i)}[X-a]\lambda_{\mathrm{m}}(B'(a)B(a)^\top+B(a)B'(a)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\quad+\lambda_{\mathrm{m}}(B(b)B(b)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\quad+\mathbb{E}_{\tilde M(i)}[(X-b)^2]\lambda_{\mathrm{m}}(B'(b)B'(b)^\top)\|\theta^0-\hat\theta\|_2^2\\
&\quad+\mathbb{E}_{\tilde M(i)}[X-b]\lambda_{\mathrm{m}}(B'(b)B(b)^\top+B(b)B'(b)^\top)\|\theta^0-\hat\theta\|_2^2.
\end{aligned}$$

Assumption (B1) imposes that $\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde M(i)}[X^2]$ and $\sup_{i\in\mathcal{I}}\lambda_{\mathrm{m}}(\mathbb{E}_{\tilde M(i)}[B(X)B(X)^\top])$ are finite. Hence, the supremum of each of the above terms is finite. That is, there exists a constant $c>0$ such that

$$\begin{aligned}
&\Big|\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde M(i)}\big[(Y-f_{\eta_0}(X;\hat\theta^n_{\lambda^\star_n,\eta_0,\mu}))^2\big]-\inf_{f_\diamond\in\mathcal{F}_{\eta_0}}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde M(i)}\big[(Y-f_\diamond(X))^2\big]\Big|\\
&\quad\le c\|\theta^0-\hat\theta^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2+2\sqrt{\mathbb{E}_M\big[\xi_Y^2\big]c}\|\theta^0-\hat\theta^n_{\lambda^\star_n,\eta_0,\mu}\|_2.
\end{aligned}$$

It therefore suffices to show that

$$\hat\theta^n_{\lambda^\star_n,\eta_0,\mu}\xrightarrow[n\to\infty]{P}\theta^0,$$

with respect to the distribution induced by M. To simplify notation, we henceforth drop the M subscript in the expectations and probabilities. Note that by the rank conditions in (B2), and the law of large numbers, we may assume that the corresponding sample product moments satisfy the same conditions. That is, for

the purpose of the following arguments, it suffices that the sample product moment only satisfies these rank conditions asymptotically with probability one.

Let $B := B(X)$, $C := C(A)$, let $\mathbf{B}$ and $\mathbf{C}$ be row-wise stacked i.i.d. copies of $B(X)^\top$ and $C(A)^\top$, and recall the definition $\mathbf{P}_\delta := \mathbf{C}\left(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M}\right)^{-1}\mathbf{C}^\top$. By convexity of the objective function we can find a closed form expression for our estimator of θ^0 by solving the corresponding normal equations. The closed form expression is given by

$$\begin{aligned}\hat{\theta}^n_{\lambda,\eta,\mu} :&= \operatorname*{arg\,min}_{\theta\in\mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma\theta^\top\mathbf{K}\theta \\ &= \left(\frac{\mathbf{B}^\top\mathbf{B}}{n} + \lambda_n^\star\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{B}}{n} + \frac{\gamma\mathbf{K}}{n}\right)^{-1}\left(\frac{\mathbf{B}^\top\mathbf{Y}}{n} + \lambda_n^\star\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{Y}}{n}\right),\end{aligned}$$

where we used that $\lambda_n^\star \in [0,\infty)$ almost surely by (C2). Consequently (using standard convergence arguments and that $n^{-1}\gamma\mathbf{K}$ and $n^{-1}\delta\mathbf{M}$ converges to zero in probability), if $\lambda_n^\star$ diverges to infinity in probability as n tends to infinity, then

$$\begin{aligned}\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu} &\xrightarrow{P} \left(\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1}\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right] \\ &= \theta^0.\end{aligned}$$

Here, we also used that the terms multiplied by $\lambda_n^\star$ are the only asymptotically relevant terms. These are the standard arguments that the K-class estimator (with minor penalized regression modifications) is consistent as long as the parameter $\lambda_n^\star$ converges to infinity, or, equivalently, $\kappa_n^\star = \lambda_n^\star/(1+\lambda_n^\star)$ converges to one in probability.

We now consider two cases: *(i)* $\mathbb{E}[B\xi_Y] \neq 0$ and *(ii)* $\mathbb{E}[B\xi_Y] = 0$, corresponding to the case with unmeasured confounding and without, respectively. For *(i)* we show that $\lambda_n^\star$ converges to infinity in probability and for *(ii)* we show consistency by other means (as $\lambda_n^\star$ might not converge to infinity in this case).

Case (i): The confounded case $\mathbb{E}[B\xi_Y] \neq 0$. It suffices to show that

$$\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \xrightarrow[n\to\infty]{P} \infty.$$

To that end, note that for fixed $\lambda \geq 0$ we have that

$$\hat{\theta}^n_{\lambda,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_\lambda, \tag{B.40}$$

where

$$\begin{aligned}\theta_\lambda :&= \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1} \\ &\quad\times\left(\mathbb{E}\left[BY\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right]\right).\end{aligned} \tag{B.41}$$

Recall that (B.39) states that $\mathbb{E}[CY] = \mathbb{E}\left[CB^\top\right]\theta^0$. Using (B.39) and that $Y = B^\top\theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that the latter factor of (B.41) is given by

$$\begin{aligned}
&\mathbb{E}[BY] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}[CY] \\
&\quad= \mathbb{E}\left[BB^\top\right]\theta^0 + \mathbb{E}[B\xi_Y] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\theta^0 \\
&\quad= \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)\theta^0 + \mathbb{E}[B\xi_Y].
\end{aligned}$$

Inserting this into (B.41) we arrive at the following representation of θ_λ

$$\theta_\lambda = \theta^0 + \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1}\mathbb{E}[B\xi_Y]. \tag{B.42}$$

Since $\mathbb{E}[B\xi_Y] \neq 0$ by assumption, the above yields that

$$\forall\lambda \geq 0: \qquad \theta^0 \neq \theta_\lambda. \tag{B.43}$$

Now we prove that $\lambda_n^\star$ diverges to infinity in probability as n tends to infinity. That is, for any $\lambda \geq 0$ we will prove that

$$\lim_{n\to\infty}\mathbb{P}(\lambda_n^\star \leq \lambda) = 0.$$

We fix an arbitrary $\lambda \geq 0$. By (B.43) we have that $\theta^0 \neq \theta_\lambda$. This implies that there exists an $\varepsilon > 0$ such that $\theta^0 \notin \overline{B(\theta_\lambda, \varepsilon)}$, where $\overline{B(\theta_\lambda, \varepsilon)}$ is the closed ball in $\mathbb{R}^k$ with center θ_λ and radius ε. By the consistency result (B.40), we know that the sequence of events $(A_n)_{n\in\mathbb{N}}$, for every $n \in \mathbb{N}$, given by

$$A_n := (|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon) = (\hat{\theta}^n_{\lambda,\eta_0,\mu} \in \overline{B(\theta_\lambda, \varepsilon)}),$$

satisfies $\mathbb{P}(A_n) \to 1$ as $n \to \infty$. By assumption (C3) we have that

$$\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu}), \qquad \text{and} \qquad \theta \mapsto T_n(\theta),$$

are weakly decreasing and continuous, respectively. Together with the continuity of $\tilde{\lambda} \mapsto \hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu}$, this implies that also the mapping $\tilde{\lambda} \mapsto T_n(\hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu})$ is continuous. It now follows from Assumption (C2) (stating that $\lambda_n^\star$ is almost surely finite) that for all $n \in \mathbb{N}$, $\mathbb{P}(T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)) = 1$. Furthermore, since $\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu})$ is weakly decreasing, it follows that

$$\begin{aligned}
\mathbb{P}(\lambda_n^\star \leq \lambda) &= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)\}) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\}) \\
&= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n) \\
&\qquad + \ \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n^c) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) \\
&\qquad + \ \mathbb{P}(A_n^c).
\end{aligned}$$

It now suffices to show that the first term converges to zero, since $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$. We have

$$\begin{aligned}
&\mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) \\
&\quad \leq \mathbb{P}\Big(\{\lambda_n^\star \leq \lambda\} \cap \Big\{ \inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}\Big) \\
&\quad \leq \mathbb{P}\Big(\inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big) \\
&\quad \xrightarrow{P} 0,
\end{aligned}$$

as $n \to \infty$, since $\overline{B(\theta_\lambda, \varepsilon)}$ is a compact set not containing θ^0. Here, we used that the test statistic (T_n) is assumed to have compact uniform power (C1). Hence, $\lim_{n\to\infty} \mathbb{P}(\lambda_n^\star \leq \lambda) = 0$ for any $\lambda \geq 0$, proving that $\lambda_n^\star$ diverges to infinity in probability, which ensures consistency.

Case (ii): the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. Recall that

$$\begin{aligned}
\hat{\theta}^n_{\lambda,\eta_0,\mu} &= \arg\min_{\theta \in \mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma \theta^\top \mathbf{K} \theta \\
&= \arg\min_{\theta \in \mathbb{R}^k} l^n_{\mathrm{OLS}}(\theta) + \lambda l^n_{\mathrm{TSLS}}(\theta) + \gamma l_{\mathrm{PEN}}(\theta),
\end{aligned} \tag{B.44}$$

where we defined $l^n_{\mathrm{OLS}}(\theta) := n^{-1}\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$, $l^n_{\mathrm{TSLS}}(\theta) := n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, and $l_{\mathrm{PEN}}(\theta) := n^{-1}\theta^\top \mathbf{K}\theta$. For any $0 \leq \lambda_1 < \lambda_2$ we have

$$\begin{aligned}
&l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_1 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \\
&\quad \leq l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_1 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad = l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_2 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\qquad + (\lambda_1 - \lambda_2) l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad \leq l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_2 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \\
&\qquad + (\lambda_1 - \lambda_2) l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),
\end{aligned}$$

where we used (B.44). Rearranging this inequality and dividing by $(\lambda_1 - \lambda_2)$ yields

$$l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \geq l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),$$

proving that $\lambda \mapsto l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda,\eta_0,\mu})$ is weakly decreasing. Thus, since $\lambda_n^\star \geq 0$ almost surely, we have that

$$l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{0,\eta_0,\mu}) = n^{-1}(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}). \tag{B.45}$$

Furthermore, recall from (B.40) that

$$\hat{\theta}^n_{0,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_0 = \theta^0, \tag{B.46}$$

where the last equality follows from (B.42) using that we are in the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. By expanding and deriving convergence statements for each term, we get

$$\begin{aligned}
&(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}) \\
&\xrightarrow[n\to\infty]{P} (\mathbb{E}[YC^\top] - \theta_0 \mathbb{E}[BC^\top])\mathbb{E}[C^\top C]^{-1}(\mathbb{E}[CY] - \mathbb{E}[CB^\top]\theta_0) \\
&= 0,
\end{aligned} \tag{B.47}$$

where we used Slutsky's theorem, the weak law of large numbers, (B.46) and (B.39). Thus, by (B.45) and (B.47) it holds that

$$l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}) = n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

For any $z \in \mathbb{R}^n$ we have that

$$\begin{aligned}
\|\mathbf{P}_\delta z\|_2^2 &= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\
&= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\
&= \|(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z\|_2^2,
\end{aligned}$$

hence

$$\begin{aligned}
\|H_n - G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2 &= \|n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \\
&\xrightarrow{P} 0,
\end{aligned} \tag{B.48}$$

where for each $n \in \mathbb{N}$, $G_n \in \mathbb{R}^{k\times k}$ and $H_n \in \mathbb{R}^{k\times 1}$ are defined as

$$\begin{aligned}
G_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{B}, \text{ and} \\
H_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{Y}.
\end{aligned}$$

Using the weak law of large numbers, the continuous mapping theorem and Slutsky's theorem, it follows that, as $n \to \infty$,

$$\begin{aligned}
G_n &\xrightarrow{P} G := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top], \text{ and} \\
H_n &\xrightarrow{P} H := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CY] \\
&\qquad = E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top]\theta^0 \\
&\qquad = G\theta^0,
\end{aligned}$$

where the second to last equality follows from (B.39). Together with (B.48), we now have that

$$\|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - G\theta^0\|_2^2 \le \|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - H_n\|_2^2 + \|H_n - G\theta^0\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

Furthermore, by the rank assumptions in (B2) we have that $G_n \in \mathbb{R}^{k\times k}$ is of full rank (with probability tending to one), hence

$$\begin{aligned}\|\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0\|_2^2 &= \|G_n^{-1}G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\leq \|G_n^{-1}\|_{\text{op}}^2\|G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\xrightarrow{P} \|G^{-1}\|_{\text{op}}^2 \cdot 0 \\ &= 0,\end{aligned}$$

as $n \to \infty$, proving the proposition.

□

Appendix C

Structure Learning For Directed Trees

C.1. Graph Terminology

A *directed graph* $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (nodes) $V = \{1, \dots, p\}$ and a collection of directed edges $\mathcal{E} \subseteq \{(j \to i) \equiv (j, i) : i, j \in V, i \neq j\}$. For any graph $\mathcal{G} = (V, \mathcal{E})$ we let $\mathrm{pa}^{\mathcal{G}}(i) := \{v \in V : \exists (v, i) \in \mathcal{E}\}$ and $\mathrm{ch}^{\mathcal{G}}(i) := \{v \in V : \exists (j, v) \in \mathcal{E}\}$ denote the *parents* and *children* of node $i \in V$ and we define root nodes $\mathrm{rt}(\mathcal{G}) := \{v \in V : \mathrm{pa}^{\mathcal{G}}(i) = \emptyset\}$ as nodes with no parents (that is, no incoming edges). A *path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2), \dots, (i_{k-1}, i_k)$ of pairs of nodes such that for all $j \in \{1, \dots, k-1\}$, we have either $(i_j \to i_{j+1}) \in \mathcal{E}$ or $(i_{j+1} \to i_j) \in \mathcal{E}$. A *directed path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2), \dots, (i_{k-1}, i_k)$ of pairs of nodes such that for all $j \in \{1, \dots, k-1\}$, we have $(i_j \to i_{j+1}) \in \mathcal{E}$. Furthermore, we let $\mathrm{an}^{\mathcal{G}}(i)$ and $\mathrm{de}^{\mathcal{G}}(i)$ denote the *ancestors* and *descendants* of node $i \in V$, consisting of all nodes $j \in V$ for which there exists a directed path to and from i, respectively. A *directed acyclic graph* (DAG) is a directed graph that does not contain any directed cycles, i.e., directed paths visiting the same node twice. We say that a graph is *connected* if a (possibly undirected) path exists between any two nodes. A *directed tree* is a connected DAG in which all nodes have at most one parent. More specifically, every node has a unique parent except the root node, which has no parent. The root node $\mathrm{rt}(\mathcal{G})$ is the unique node such there exists a directed path from $\mathrm{rt}(\mathcal{G})$ to any other node in the directed tree. In graph theory, a directed tree is also called an *arborescence*, a *directed rooted tree*, and a *rooted out-tree*. A graph $\mathcal{G} = (V', \mathcal{E}')$ is a *subgraph* of another graph $\mathcal{G} = (V, \mathcal{E})$ if $V' \subseteq V$, $\mathcal{E}' \subseteq \mathcal{E}$ and for all $(j \to i) \in \mathcal{E}'$ it holds that $j, i \in V'$. A subgraph is *spanning* if $V' = V$.

C.2. Further Details on Section 4.5

Remark C.1. The conditional entropy score gap is not strictly positive when considering the alternative graphs $\tilde{\mathcal{G}}$ that are Markov equivalent to the causal graph $\mathcal{G}$, $\tilde{\mathcal{G}} \in \mathrm{MEC}(\mathcal{G})$. A simple translation of the conditional entropy score function reveals that

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) + C = \sum_{(j \to i) \in \tilde{\mathcal{E}}} h(X_i|X_j) - h(X_i) = - \sum_{(j \to i) \in \tilde{\mathcal{E}}} I(X_i; X_j),$$

for a constant $C \in \mathbb{R}$. By symmetry of the mutual information, it holds that $\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) = \ell_{\mathrm{CE}}(\mathcal{G})$, for any $\tilde{\mathcal{G}} \in \mathrm{MEC}(\mathcal{G})$, since $\tilde{\mathcal{G}}$ and $\mathcal{G}$ share the same skeleton. Thus, the conditional entropy score function can, at most, identify the Markov equivalence class of the causal graph. In fact, the polytree causal structure learning method of Rebane and Pearl (1987) uses the above translated conditional entropy score function to recover the skeleton of the causal graph. ∘

Example C.1 (Negative local Gaussian score gap)**.** Consider two graphs $\mathcal{G}$ and $\tilde{\mathcal{G}}$ with different root nodes, i.e., $\mathrm{rt}(\mathcal{G}) \neq \mathrm{rt}(\tilde{\mathcal{G}})$. If $x \mapsto \mathbb{E}[X_{\mathrm{rt}(\mathcal{G})}|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))} = x]$ is not almost surely constant, then it holds that

$$\begin{aligned}
\ell_{\mathrm{G}}(\tilde{\mathcal{G}}, \mathrm{rt}(\mathcal{G})) - \ell_{\mathrm{G}}(\mathcal{G}, \mathrm{rt}(\mathcal{G})) &= \mathbb{E}[(X_{\mathrm{rt}(\mathcal{G})} - \mathbb{E}[X_{\mathrm{rt}(\mathcal{G})}|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))}])^2] \\
&\quad - \mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})}) \\
&= \mathbb{E}[\mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})}|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))})] - \mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})}) \\
&= -\mathrm{Var}(\mathbb{E}[X_{\mathrm{rt}(\mathcal{G})}|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))}]) \\
&< 0.
\end{aligned}$$

∘

C.3. Further Details on the Simulation Experiments

This section contains further details on the simulation experiments.

C.3.1. Tree Generation Algorithms

The following two algorithms, Algorithm C.1 (many leaf nodes) and Algorithm C.2 (many branch nodes), details how the Type 1 and Type 2 trees are generated, respectively.

C.3.2. Additional Illustrations

This section contains some additional illustrations of the simulation experiments.

Algorithm C.1 Generating type 1 trees

procedure TYPE1(p)
  $A := 0 \in \mathbb{R}^{p \times p}$
  for $j \in \{1, \ldots, p\}$ **do**
    for $i \in \{j + 1, \ldots, p\}$ **do**
      if $\sum_{k=1}^{p} A_{ki} = 0$ **then**
        if $i = j + 1$ **then**
          $A_{ji} := 1$
        else
          $A_{ji} :=$ Binomial(success $= 0.1$)
        end if
      else
        $A_{ji} := 0$
      end if
    end for
  end for
  return A
end procedure

Algorithm C.2 Generating type 2 trees

procedure TYPE2(p)
  for $i \in \{2, \ldots, p\}$ **do**
    $j :=$ sample($\{1, \ldots, i - 1\}$)
    $A_{ji} := 1$
  end for
  return A
end procedure

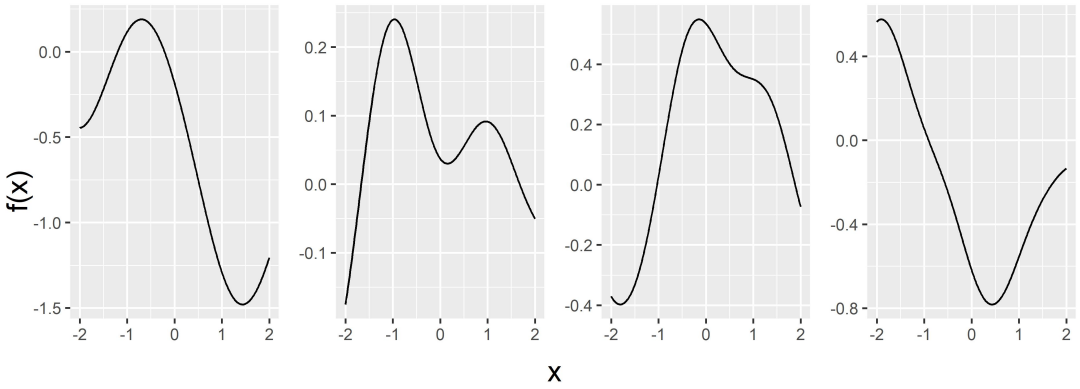


Figure C.1: Four causal functions as modeled by the RBF kernel Gaussian Process.

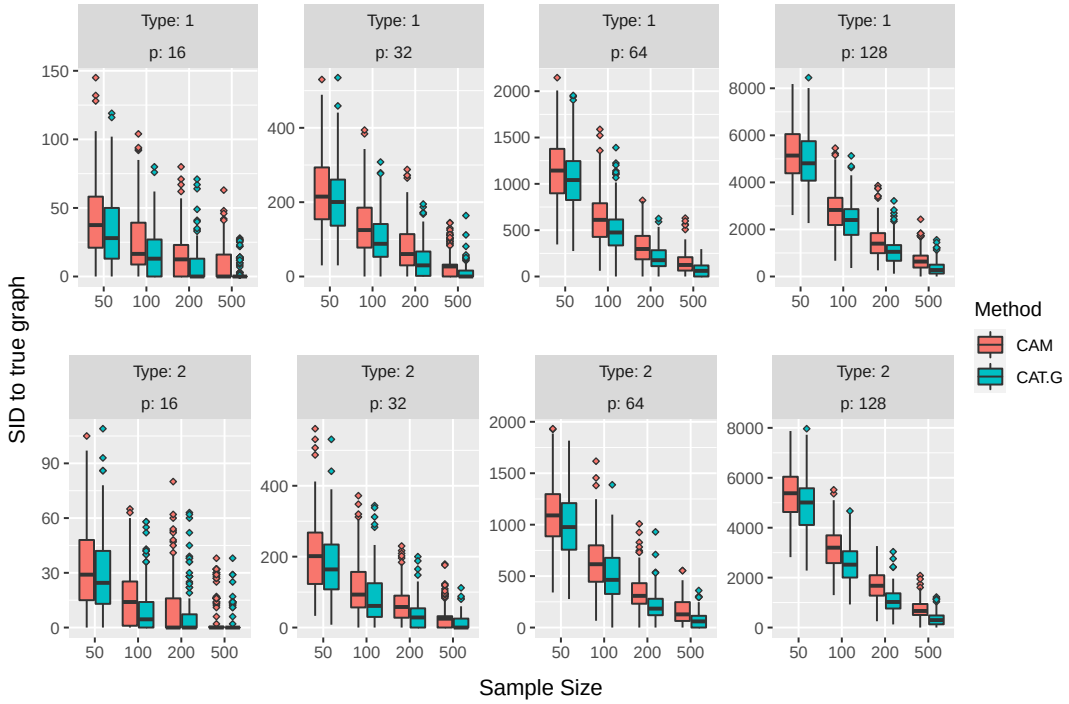


Figure C.2: Boxplot illustrating the SID performance of CAM and CAT for varying sample sizes, system sizes and tree types in the experiment of Section 4.6.1.2. CAT.G is CAT with edge weights derived from the Gaussian score function.

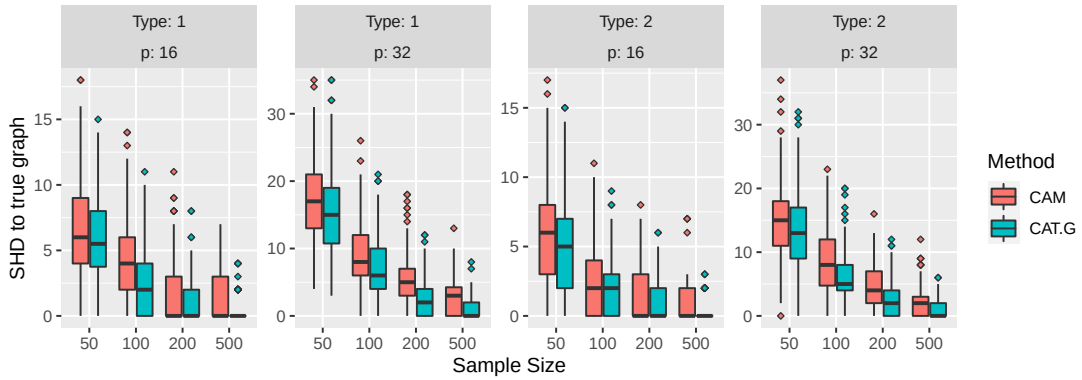


Figure C.3: Boxplot illustrating the SHD performance of CAM and CAT for varying sample sizes, system sizes and tree types in the experiment of Section 4.6.1.2. Here CAT.G is run on the CAM edge weights , so that any difference in nonparametric regression technique is ruled out as the source of the performance difference.

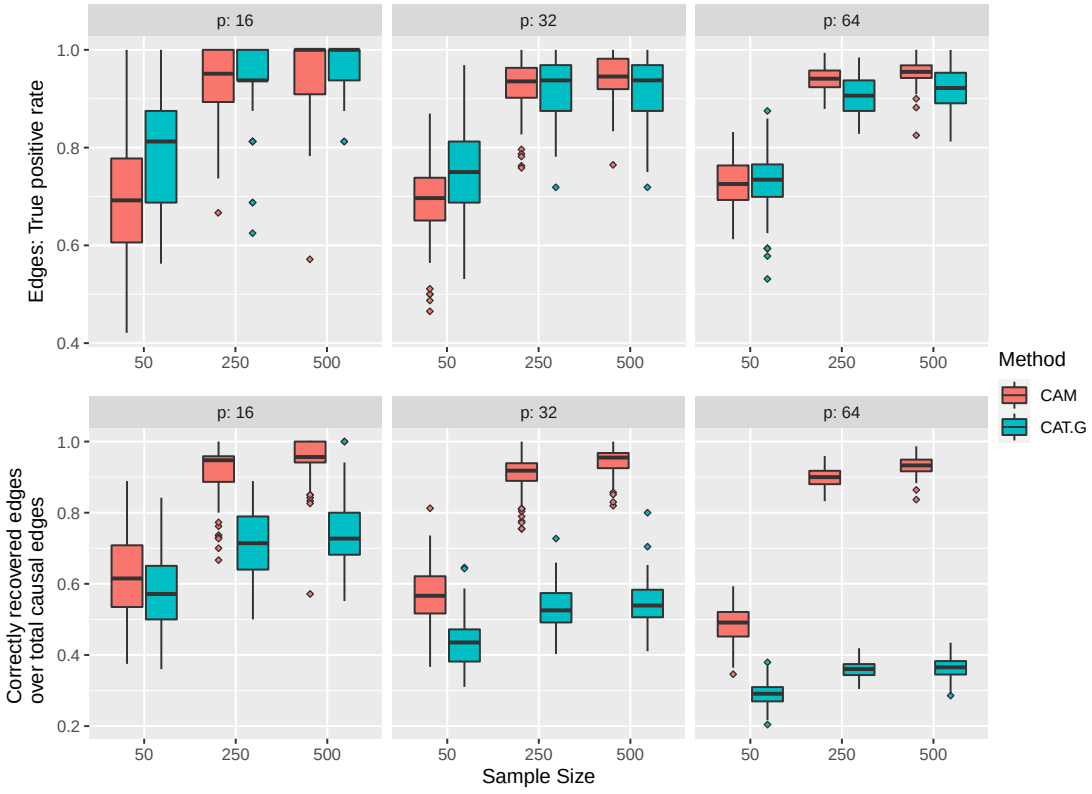


Figure C.4: Boxplot of edge relations for the experiment in Section 4.6.3.

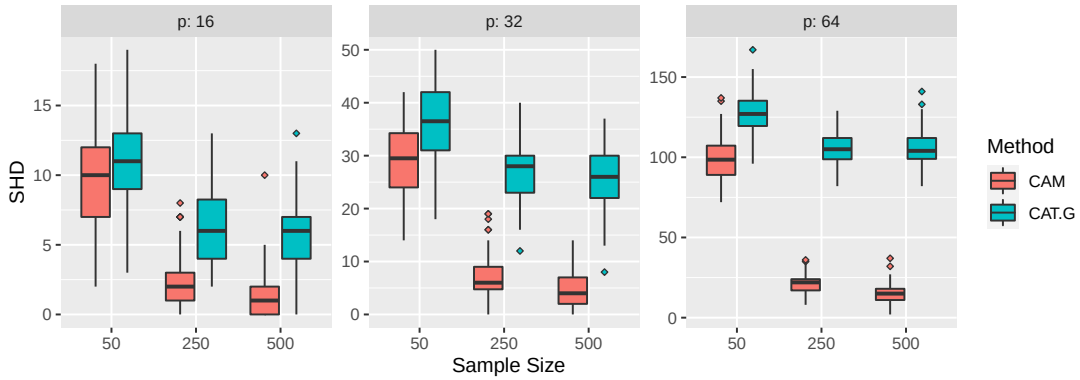


Figure C.5: Boxplot of SHD for the experiment in Section 4.6.3.

C.4. Proofs

This section contains the proofs of all results presented in the main text.

C.4.1. Proofs of Section 4.2

Proof of Lemma 4.1: Fix $i \in \{1, \dots, p\} \setminus \mathrm{rt}(\mathcal{G})$ and assume that $P_N \in \mathcal{P}_{\mathrm{G}}^p$. Furthermore, let $f_i \in \mathcal{D}_3$ and nowhere constant and nonlinear. Assume for contradiction that each f_i satisfy does not satisfy Equation (4.1). Recall that the additive noise is Gaussian, so

$$\nu(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}, \quad \nu'(x) = -\frac{x}{\sigma^2}, \quad \nu''(x) = -\frac{1}{\sigma^2}, \quad \nu'''(x) = 0.$$

Hence, the negation of Equation (4.1) reduces to

$$\xi'''(x) - \xi''(x)\frac{f''(x)}{f'(x)} - \frac{2f''(x)f'(x)}{\sigma^2} = -\frac{y - f(x)}{\sigma^2}\left(f'''(x) - \frac{(f''(x))^2}{f'(x)}\right), \quad \text{(C.1)}$$

for any $(x, y) \in \mathcal{J} = \{(x, y) \in \mathbb{R}^2 : f'(x) \neq 0\}$. Furthermore, f is nowhere constant, so $f'(x) = 0$ for at most countably many $x \in \mathbb{R}$. As the left-hand side of Equation (C.1) is constant in y it must hold that

$$0 = f'''(x) - \frac{(f''(x))^2}{f'(x)} = \frac{\frac{\partial f''(x)}{\partial x}f'(x) - f''(x)\frac{\partial f'(x)}{\partial x}}{(f'(x))^2} = \frac{\partial}{\partial x}\left(\frac{f''(x)}{f'(x)}\right),$$

i.e., $f''(x)/f'(x)$ is constant for all $x \in \mathbb{R}$ such that $f'(x) \neq 0$. Now note that there exists a countable collection of disjoint open intervals (O_k) that covers $\mathbb{R}$ almost everywhere such $f'(x) \neq 0$ on each O_k. Assume for contradiction that $f''(x)/f'(x) = c_{k,1} \neq 0$ on some O_k. On each O_k we have that $\partial/\partial x \log(\mathrm{sign}(f'(x))f'(x)) = c_{k,1} \iff \log(\mathrm{sign}(f'(x))f'(x)) = c_{k,1}x + c_{k,2} \iff \mathrm{sign}(f'(x))f'(x) = \exp(c_{k,1}x + c_{k,2}) \iff f'(x) = \pm\exp(c_{k,1}x + c_{k,2})$. Assume without loss of generality that $f'(x) = \exp(c_{k,1}x + c_{k,2})$ for all $x \in O_k$ and k. By the assumed continuous differentiability of f' we need to stitch these functions together in a continuously differentiable way. That is, we require that for any k that $t_k = \sup(O_k) = \inf(O_{k+1})$ and $\lim_{x\uparrow t_k} f'(x) = \lim_{x\downarrow t_k} f'(x)$ and $\lim_{x\uparrow t_k} f''(x) = \lim_{x\downarrow t_k} f''(x)$. These conditions impose the restrictions $(c_{k,1} - c_{k+1,1})t_k = c_{k+1,2} - c_{k,2}$ and $\log(c_{k,1}/c_{k+1,1}) + (c_{k,1} - c_{k+1,1})t_k = c_{k+1,2} - c_{k,2}$ which entails that $c_{k,1} = c_{k+1,1}$ and $c_{k,2} = c_{k+1,2}$. This proves that there exists $c_1, c_2 \in \mathbb{R}$ such that $f'(x) = \pm\exp(c_1x + c_2)$ for all $x \in \mathbb{R}$. Thus, the differential equation holds for all $x \in \mathbb{R}$,

$$0 = \xi'''(x) - \xi''(x)\frac{f''(x)}{f'(x)} - \frac{2f''(x)f'(x)}{\sigma^2} = \frac{\partial}{\partial x}\left(\frac{\xi''(x)}{f'(x)}\right) - 2\frac{f''(x)}{\sigma^2},$$

by division with $f'(x)$. By integration this implies that $0 = \xi''(x)/f'(x) - 2f'(x)/\sigma^2 + c_3$ such that $\xi''(x) = 2\exp(2c_1x + 2c_2)/\sigma^2 - c_3\exp(c_1x + c_2)$ and $\xi'(x) = \exp(2c_1x + 2c_2)/c_1\sigma^2 - c_3\exp(c_1x + c_2)/c_1 + c_4$ and

$$\xi(x) = \frac{\exp(2c_1x + 2c_2)}{2c_1^2\sigma^2} - \frac{c_3\exp(c_1x + c_2)}{c_1^2} + c_4x + c_5.$$

We see that $\xi(x) \to \infty \iff p_X(x) \to \infty$ as $x \to \operatorname{sign}(c_1) \cdot \infty$, in contradiction with the assumption that p_X is a probability density function. Thus, it must hold that $f''(x)/f'(x) = 0$ or equivalently that f is a linear function, a contradiction. This proves that whenever $f_i \in \mathcal{D}_3$ is a nowhere constant and nonlinear function and the additive noise is Gaussian then Equation (4.1) holds. □

Proof of Proposition 4.1: First, we consider the bivariate setting. Let (X, Y) be generated by an additive noise SCM $\theta \in \Theta_R \subseteq \mathcal{T}_2 \times \mathcal{D}_3^2 \times \mathcal{P}_{\mathcal{C}_3}^2$ given by $X := N_X$ and $Y := f(X) + N_Y$ with $P_X = p_X \cdot \lambda$ and $P_{N_Y} = p_{N_Y} \cdot \lambda$ having three times differentiable strictly positive densities and f is a three times differentiable nowhere constant function such that Equation (4.1) holds.

Assume for contradiction that we do not have observational identifiability of the causal structure $\mathcal{G} = (V = \{X, Y\}, \mathcal{E} = \{(X \to Y)\})$. That is, there exists $\tilde{\theta} \in \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$ with causal graph $\tilde{\mathcal{G}} \neq \mathcal{G}$ or, equivalently, a differentiable function g and noise distributions $P_{\tilde{N}_X} = p_{\tilde{N}_X} \cdot \lambda$ and $P_{\tilde{N}_Y} = p_{\tilde{N}_Y} \cdot \lambda$ with continuous densities such that structural assignments $\tilde{Y} := \tilde{N}_Y$ and $\tilde{X} := g(\tilde{Y}) + \tilde{N}_X$ induces an identical distribution, i.e., that

$$P_{X,Y} = P_{\tilde{X},\tilde{Y}}. \tag{C.2}$$

By the additive noise structural assignments we know that both $P_{X,Y}$ and $P_{\tilde{X},\tilde{Y}}$ have densities with respect to λ^2 given by

$$p_{X,Y}(x, y) = p_X(x)p_{N_Y}(y - f(x)),$$
$$p_{\tilde{X},\tilde{Y}}(x, y) = p_{\tilde{N}_X}(x - g(y))p_{\tilde{Y}}(y),$$

for all $(x, y) \in \mathbb{R}^2$. By the equality of distributions in Equation (C.2) and strict positivity of p_X and p_{N_Y} we especially have that for λ^2-almost all $(x, y) \in \mathbb{R}^2$

$$0 < p_{X,Y}(x, y) = p_{\tilde{X},\tilde{Y}}(x, y). \tag{C.3}$$

However, as both $p_{X,Y}$ and $p_{\tilde{X},\tilde{Y}}$ are continuous we realize that Equation (C.3) holds for all $(x, y) \in \mathbb{R}$. If they were not everywhere equal there would exists a non-empty open ball in $\mathbb{R}^2$ on which they differ in contradiction with λ^2-almost everywhere equality. Furthermore, by the assumption that f is three times differentiable and p_X, p_{N_Y} are three times continuously differentiable we have that $\partial^3\pi/\partial x^3$ and $\partial^3\pi/\partial x^2\partial y$ are well-defined partial-derivatives of $\pi(x, y) := \log p_{X,Y}(x, y)$ given by

$$\pi(x, y) = \log p_X(x) + \log p_{N_Y}(y - f(x)) =: \xi(x) + \nu(y - f(x)),$$

With $\tilde{\pi}(x,y) := \log p_{\tilde{X},\tilde{Y}}$ we have that

$$\tilde{\pi}(x,y) = \log p_{\tilde{N}_X}(x - g(y)) + \log p_{\tilde{Y}}(y) =: \tilde{\xi}(x - g(y)) + \tilde{\nu}(y).$$

Since it holds that $\pi = \tilde{\pi}$ by Equation (C.3) the partial-derivatives $\partial^3\tilde{\pi}/\partial x^3$ and $\partial^3\tilde{\pi}/\partial x^2\partial y$ are also well-defined. Now note that for any $x, y \in \mathbb{R}$

$$0 = \lim_{h\to 0} |\tilde{\pi}(x+h,y) - \tilde{\pi}(x,y)|/h = \lim_{h\to 0} |\tilde{\xi}(x - g(y) + h) - \tilde{\xi}(x - g(y))|/h,$$

implying that $\tilde{\xi}$ is differentiable in $x - g(y)$ for any $x, y \in \mathbb{R}$ or, equivalently, $\tilde{\xi}$ is everywhere differentiable. Similar arguments yield that $\tilde{\xi}$ is at least three times differentiable. We conclude that $\partial^2\tilde{\pi}(x,y)/\partial x^2 = \tilde{\xi}''(x - g(y))$ and $\partial^2\tilde{\pi}(x,y)/\partial x\partial y = -\tilde{\xi}''(x - g(y))g'(y)$ and for any $(x,y) \in \mathbb{R}^2$ such that $\partial^2\tilde{\pi}(x,y)/\partial x\partial y \neq 0$ or, equivalently,

$$\forall (x,y) \in \mathcal{J} := \left\{(x,y) : \frac{\partial^2\pi(x,y)}{\partial x\partial y} = -\nu''(y - f(x))f'(x) \neq 0\right\},$$

it holds that

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2}{\partial x^2}\tilde{\pi}(x,y)}{\frac{\partial^2}{\partial x\partial y}\tilde{\pi}(x,y)}\right) = \frac{\partial}{\partial x}\left(\frac{-1}{g'(y)}\right) = 0.$$

It is worth noting that $\mathcal{J} \neq \emptyset$ to ensure that the following derivations are not void of meaning. This can be seen by noting that f is nowhere constant, i.e., $f'(x) \neq 0$ for λ-almost all $x \in \mathbb{R}$. Hence, $\mathcal{J} = \emptyset$ if and only if p_{N_Y} is a density such that $\{(x,y) \in \mathbb{R}^2 : f'(x) \neq 0\} \ni (x,y) \mapsto \nu''(y - f(x))$ is constantly zero or equivalently $\mathbb{R} \ni y \mapsto \nu''(y)$ is constantly zero. This holds if and only if p_{N_Y} is either exponentially decreasing or exponentially increasing everywhere, which is a contradiction as no continuously differentiable function integrating to one has this property. On the other hand, for any $(x,y) \in \mathcal{J}$ we also have that

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2}{\partial x^2}\pi(x,y)}{\frac{\partial^2}{\partial x\partial y}\pi(x,y)}\right) \\
&= \frac{\partial}{\partial x}\left(\frac{\xi''(x) + \nu''(y - f(x))f'(x)^2 - \nu'(y - f(x))f''(x)}{-\nu''(y - f(x))f'(x)}\right) \\
&= -2f'' + \frac{\nu' f'''}{\nu'' f'} - \frac{\xi'''}{\nu'' f'} + \frac{\nu'''\nu' f''}{(\nu'')^2} \\
&\quad - \frac{\nu'''\xi''}{(\nu'')^2} - \frac{(f'')^2\nu'}{\nu''(f')^2} + \frac{f''\xi''}{\nu''(f')^2},
\end{aligned}$$

which implies that

$$\xi''' = \xi''\left(\frac{f''}{f'} - \frac{f'\nu'''}{\nu''}\right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu'''\nu' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}, \tag{C.4}$$

in contradiction with the assumption that Equation (4.1) holds. We conclude that $P_{X,Y} \neq P_{\tilde{X},\tilde{Y}}$.

Now consider a multivariate restricted causal model over $X = (X_1, \dots, X_p)$ with causal directed tree graph $\mathcal{G} = (V, \mathcal{E})$. Assume for contradiction that there exists an alternative model $\tilde{\theta} \in \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$ inducing $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_p)$ with causal graph $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \neq \mathcal{G}$, such that $P_X = P_{\tilde{X}}$. As P_X is Markovian with respect to both $\mathcal{G}$ and $\tilde{\mathcal{G}}$, i.e., the graphs are Markov equivalent. We know by Lemma C.6 that there exists a directed path in $\mathcal{G}$ that is reversed in $\tilde{\mathcal{G}}$. We especially have that there exists $i, j \in V$ such that $(j \to i) \in \mathcal{E}$ and $(i \to j) \in \tilde{\mathcal{E}}$. That is, the following structural equations hold for (X_i, X_j) and $(\tilde{X}_i, \tilde{X}_j)$

$$X_i = f_i(X_j) + N_i, \quad \text{with} \quad X_j \perp\!\!\!\perp N_i, \tag{C.5}$$

$$\tilde{X}_j = \tilde{f}_j(\tilde{X}_i) + \tilde{N}_j, \quad \text{with} \quad \tilde{X}_i \perp\!\!\!\perp \tilde{N}_j, \tag{C.6}$$

with $P_{X_j,X_i} = P_{\tilde{X}_j,\tilde{X}_i}$. We can apply the exact same arguments as in the bivariate setup if we can argue that p_{X_j} is three times differentiable and that $p_{\tilde{X}_i}$ is a continuous density.

To this end, note that the density p_{X_j} is given by the convolution of two densities

$$p_{X_j}(y) = \int_{-\infty}^{\infty} p_{f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)})}(t) p_{N_j}(y - t)\, dt, \tag{C.7}$$

as $X_j := f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)}) + N_j$ with $X_{\mathrm{pa}^{\mathcal{G}}(j)} \perp\!\!\!\perp N_j$. Here we used that $f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)})$ has density with respect to the Lebesgue measure. To realize this note that $f_j \in \mathcal{C}_3$ and it is nowhere constant. This implies that $f'(x) = 0$ at only countably many points (d_k). Now let (O_k) be the collection of disjoint open intervals that cover $\mathbb{R}$ except for the points (d_k). By continuity of f' we know that $f'(x)$ is either strictly positive or strictly negative on each O_k. That is, f is continuously differentiable and strictly monotone on each O_k. Thus, f has a continuously differentiable inverse on each O_k by, e.g., the inverse function theorem. This ensures that $f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)})$ has density with respect to the Lebesgue measure whenever $X_{\mathrm{pa}^{\mathcal{G}}(j)}$ does. By starting at the root node $X_{\mathrm{rt}(\mathcal{G})} = N_{\mathrm{rt}(\mathcal{G})}$ which by assumption has density, we can iteratively apply the above argumentation down the directed path from $\mathrm{rt}(\mathcal{G})$ to j in order to conclude that any X_j for $j \in \{1, \dots, p\}$ has density with respect to the Lebesgue measure. Since p_{N_j} is assumed strictly positive three times continuous differentiable, the representation in Equation (C.5) furthermore yields that p_{X_j} is three times differentiable; see, e.g., Theorem 11.4 and 11.5 of Schilling (2017).

Now we argue that $p_{\tilde{X}_i}$ is continuous, or more specifically that it has a continuous version. First note that P_{X_i} at least has a continuous density p_{X_i} by arguments similar to those applied for Equation (C.7). By the assumption that $P_X = P_{\tilde{X}}$ we especially have that $P_{X_i} = P_{\tilde{X}_i}$ which implies that also $\tilde{X}_i$ has a continuous density. By virtue of the arguments for the bivariate setup we arrive at a contradiction, so it must hold that $P_X \neq P_{\tilde{X}}$. □

Proof of Lemma 4.2: Consider any SCM $\tilde{\theta} = (\tilde{\mathcal{G}}, (\tilde{f}_i), P_{\tilde{N}}) \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p$ with $\tilde{\mathcal{G}} \neq \mathcal{G}$ and let $Q_{\tilde{\theta}}$ be the induced distribution. As $Q_{\tilde{\theta}}$ is Markov with respect

to $\tilde{\mathcal{G}}$ and generated by an additive noise model the density $q_{\tilde{\theta}}$ factorizes as

$$q_{\tilde{\theta}}(x) = \prod_{i=1}^{p} q_{\tilde{\theta}}(x_i | x_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) = \prod_{i=1}^{p} q_{\tilde{N}_i}(x_i - \tilde{f}_i(x_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})).$$

The cross entropy between P_X and $Q_{\tilde{\theta}}$ is then given by

$$\begin{aligned} h(P_X, Q_{\tilde{\theta}}) &= \mathbb{E}\left[-\log\left(q_{\tilde{\theta}}(X)\right)\right] \\ &= \sum_{i=1}^{p} \mathbb{E}\left[-\log\left(q_{\tilde{N}_i}\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)\right)\right] \\ &= \sum_{i=1}^{p} h\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}), \tilde{N}_i\right). \end{aligned}$$

As $Q_{\tilde{\theta}}$ is generated by a Gaussian noise structural causal model, we know that $\tilde{N}_i \sim \mathcal{N}(0, \tilde{\sigma}_i^2)$ for some $\tilde{\sigma}_i^2 > 0$ for all $1 \leq i \leq p$. Hence

$$\begin{aligned} & h\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}), \tilde{N}_i\right) \\ &= \mathbb{E}\left[-\log\left(\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2}{2\tilde{\sigma}_i^2}\right)\right)\right] \\ &= \log(\sqrt{2\pi}\tilde{\sigma}_i) + \frac{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]}{2\tilde{\sigma}_i^2}, \end{aligned}$$

for all $1 \leq i \leq p$. Thus, for given set of causal functions $(\tilde{f}_i)$ we get that the noise variances that minimizes the term-wise cross entropy is given by

$$\tilde{\sigma}_i = \sqrt{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]},$$

and attains the value

$$\begin{aligned} & \inf_{\tilde{\sigma}_i > 0}\left\{\log(\sqrt{2\pi}\tilde{\sigma}_i) + \frac{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]}{2\tilde{\sigma}_i^2}\right\} \\ &= \log\left(\sqrt{2\pi}\right) + \frac{1}{2}\log\left(\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]\right) + \frac{1}{2}. \end{aligned}$$

We conclude that

$$\begin{aligned} & \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) \\ &= p\log(\sqrt{2\pi}) + \frac{p}{2} + \sum_{i=1}^{p} \frac{1}{2}\log\left(\inf_{\tilde{f}_i \in \mathcal{D}_1} \mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]\right). \end{aligned}$$

Finally, note that as $\mathcal{D}_1$ is dense in $\mathcal{L}^2(P_{X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}})$, we have that

$$\begin{aligned}\inf_{\tilde{f}_i \in \mathcal{D}_1} \mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})\right)^2\right] &= \mathbb{E}\left[\left(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}])\right)^2\right] \\ &\quad + \inf_{\tilde{f}_i \in \mathcal{D}_1} \mathbb{E}\left[\left(\mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}] - \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})\right)^2\right] \\ &= \mathbb{E}\left[\left(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}])\right)^2\right].\end{aligned}$$

Here we used that $X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}$ has density with respect to the Lebesgue measure, $P_{X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}} \ll \lambda$, and that the density is differentiable (see proof of Proposition 4.1). This concludes the first part of the proof.

For the second statement, we note that for any $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ there exists some noise innovation distribution $P_{\tilde{N}} \in \mathcal{P}$ such that Q is the distribution of $\tilde{X}$ generated by structural assignments

$$\tilde{X}_i := \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}) + \tilde{N}_i = \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}] + \tilde{N}_i,$$

for all $1 \leq j \leq p$ and mutually independent noise innovations $\tilde{N} = (\tilde{N}_1, \ldots, \tilde{N}_p) \sim P_{\tilde{N}} \in \mathcal{P}^p$. Let q denote the density of Q with respect to the Lebesgue measure λ and let $q_{\tilde{N}_i}$ denote the density of $\tilde{N}_i$ for all $1 \leq i \leq p$. As Q is Markov with respect to $\tilde{\mathcal{G}}$ and generated by an additive noise model the density factorizes as

$$q(x) = \prod_{i=1}^p q(x_i|x_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}) = \prod_{i=1}^p q_{\tilde{N}_i}(x_i - \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)} = x_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}]).$$

The cross entropy between P_X and Q is given by

$$\begin{aligned}h(P_X, Q) &= \mathbb{E}\left[-\log\left(q(X)\right)\right] \\ &= \sum_{i=1}^p \mathbb{E}\left[-\log\left(q(X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})\right)\right] \\ &= \sum_{i=1}^p \mathbb{E}\left[-\log\left(q_{\tilde{N}_i}\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right]\right)\right)\right] \\ &= \sum_{i=1}^p h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right).\end{aligned}$$

Note that $h(P, Q) = h(P) + D_{\mathrm{KL}}(P\|Q) \geq h(P)$ with equality if and only if $Q = P$. Thus, the infimum is attained at noise innovations that are equal in distribution to $X_i - \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}]$ (which has a density by assumption). That is,

$$\begin{aligned}\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) &= \sum_{i=1}^p \inf_{\tilde{N}_j \sim P_{\tilde{N}_j} \in \mathcal{P}} h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\ &= \sum_{i=1}^p h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right]\right) \\ &= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}).\end{aligned}$$

□

C.4.2. Proofs of Section 4.3

Proof of Theorem 4.1: For simplicity of the proof we assume that $\mathbb{E}[X] = 0$ such that the edge weight estimators simplify to

$$\hat{w}_{\mathrm{G}}(j \to i) := \hat{w}_{ji}(\mathbf{X}_n, \tilde{\mathbf{X}}_n) := \frac{1}{2} \log \left(\frac{\frac{1}{n} \sum_{k=1}^{n} \left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}) \right)^2}{\frac{1}{n} \sum_{k=1}^{n} X_{k,i}^2} \right).$$

We assume that $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R$ with $\mathcal{G} = (V, \mathcal{E})$, which by Assumption 4.1 implies that there exists an $m > 0$ such that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = m > 0. \tag{C.8}$$

Let l^*_{G} auxiliary population score function such that $\ell^*_{\mathrm{G}}(\tilde{\mathcal{G}}) \geq \ell_{\mathrm{G}}(\tilde{\mathcal{G}})$ for all $\tilde{\mathcal{G}} \in \mathcal{T}_p$. For any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$ it holds that

$$\ell_{\mathrm{G}}(\mathcal{G}) + \frac{m}{2} \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \frac{m}{2} \leq \ell^*_{\mathrm{G}}(\tilde{\mathcal{G}}) - \frac{m}{2},$$

by the identifiability assumption of Equation (C.8). Thus, we have that

$$\begin{aligned} &P\left(\arg\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \hat{l}_n(\tilde{\mathcal{G}}) = \mathcal{G} \right) \\ &\geq P\left(\left(|\hat{\ell}_{\mathrm{G}}(\mathcal{G}) - \ell_{\mathrm{G}}(\mathcal{G})| < \frac{m}{2} \right) \cap \bigcap_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \left(|\hat{\ell}_{\mathrm{G}}(\tilde{\mathcal{G}}) - l^*_{\mathrm{G}}(\tilde{\mathcal{G}})| < \frac{m}{2} \right) \right), \end{aligned}$$

so it suffices to show that

$$\hat{l}_{\mathrm{G}}(\mathcal{G}) \xrightarrow{P} \ell_{\mathrm{G}}(\mathcal{G}) \quad \text{and} \quad \forall \tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \hat{\ell}_{\mathrm{G}}(\tilde{\mathcal{G}}) \xrightarrow{P} \ell^*_{\mathrm{G}}(\tilde{\mathcal{G}}), \qquad \text{as } n \to \infty.$$

We let $\ell^*_{\mathrm{G}} : \mathcal{T}_p \to \mathbb{R}$ be given by

$$\begin{aligned} l^*_{\mathrm{G}}(\tilde{\mathcal{G}}) = &\sum_{(j \to i) \in \tilde{\mathcal{E}} \setminus \mathcal{E}} \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \\ &+ \sum_{(j \to i) \in \tilde{\mathcal{E}} \cap \mathcal{E}} \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right), \end{aligned}$$

for any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$. As the conditional expectation minimizes the MSPE among measurable functions, i.e., $\varphi_{i,j} = \arg\min_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(X_i - f(X_j))^2]$, we especially have that $\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2] \geq \mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]$ for any $i, j \in V$ with $i \neq j$. This construction entails that both $l^*_{\mathrm{G}}(\tilde{\mathcal{G}}) \geq \ell_{\mathrm{G}}(\tilde{\mathcal{G}})$ for any $\tilde{\mathcal{G}} \in \mathcal{T}_p$ and that $l^*_{\mathrm{G}}(\mathcal{G}) = \ell_{\mathrm{G}}(\mathcal{G})$. We conclude that it suffices to show that

$$\sup_{\tilde{\mathcal{G}} \in \mathcal{T}_p} |\hat{\ell}_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell^*_{\mathrm{G}}(\tilde{\mathcal{G}})| \xrightarrow{P}_n 0.$$

To this end, let $\mathcal{E}^* = \{(j \to i) : i, j \in V, i \neq j\} \setminus \mathcal{E}$ and note that

$$
\begin{aligned}
&\sup_{\tilde{\mathcal{G}} \in \mathcal{T}_p} |\hat{\ell}_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell^*_{\mathrm{G}}(\tilde{\mathcal{G}})| \\
&\leq \sup_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \Bigg(\sum_{(j \to i) \in \tilde{\mathcal{E}} \setminus \mathcal{E}} \left| \hat{w}_{\mathrm{G}}(j \to i) - \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \right| \\
&\qquad + \sum_{(j \to i) \in \tilde{\mathcal{E}} \cap \mathcal{E}} \left| \hat{w}_{\mathrm{G}}(j \to i) - \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \right| \Bigg) \\
&\leq \sum_{(j \to i) \in \mathcal{E}^*} \left| \hat{w}_{\mathrm{G}}(j \to i) - \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \right| \\
&\qquad + \sum_{(j \to i) \in \mathcal{E}} \left| \hat{w}_{\mathrm{G}}(j \to i) - \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \right| . \qquad \text{(C.9)}
\end{aligned}
$$

Now consider a fixed term $(j \to i) \in \mathcal{E}$ in the second sum of Equation (C.9). We can upper bound the difference by

$$
\begin{aligned}
&\left| \hat{w}_{\mathrm{G}}(j \to i) - \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right) \right| \\
&\leq \frac{1}{2} \left| \log \left(\frac{1}{n} \sum_{k=1}^{n} \left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}) \right)^2 \right) - \log \left(\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2] \right) \right| \\
&\quad + \frac{1}{2} \left| \log(\mathbb{E}[X_i^2]) - \log \left(\frac{1}{n} \sum_{k=1}^{n} X_{k,i}^2 \right) \right| . \qquad \text{(C.10)}
\end{aligned}
$$

In the upper bound of Equation (C.10) the last two terms vanish in probability due to the law of large numbers and the continuous mapping theorem. The two first terms also vanishes by the following arguments:

$$
\begin{aligned}
0 &\leq \frac{1}{n} \sum_{k=1}^{n} \left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}) \right)^2 \\
&= \frac{1}{n} \sum_{k=1}^{n} \left(X_{k,i} - \varphi_{ji}(X_{k,j}) \right)^2 + \frac{1}{n} \sum_{k=1}^{n} \left(\varphi_{ji}(X_{k,j}) - \hat{\varphi}_{ji}(X_{k,j}) \right)^2 \\
&\quad + \frac{2}{n} \sum_{k=1}^{n} \left(X_{k,i} - \varphi_{ji}(X_{k,j}) \right) \left(\varphi_{ji}(X_{k,j}) - \hat{\varphi}_{ji}(X_{k,j}) \right) .
\end{aligned}
$$

Hence, it holds that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i}-\hat{\varphi}_{ji}(X_{k,j})\right)^2-\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2\right| \\
=&\left|\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right. \\
&\quad\left.+\frac{2}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)\right| \\
\leq&\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2 \\
&\quad+2\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2}\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2}, \qquad \text{(C.11)}
\end{aligned}$$

by Cauchy-Schwarz inequality. By the law of large numbers, we have that the first factor of the second term of Equation (C.11) converges in probability to a constant,

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2 \xrightarrow{P}_n \mathbb{E}[X_{1,i}-\varphi_{i,j}(X_{1,j}))^2].$$

The first term and latter factor of the second term of Equation (C.11) vanishes in probability by assumption. That is, for any $\varepsilon > 0$ we have that

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{k=1}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right|>\varepsilon\right) \\
&=P\left(\left|\frac{1}{n}\sum_{k=1}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right|\wedge\varepsilon>\varepsilon\right) \\
&\leq\frac{\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right)\wedge\varepsilon\right]}{\varepsilon} \\
&\leq\frac{\mathbb{E}\left[\mathbb{E}\left[\left(\varphi_{ji}(X_{1,j})-\hat{\varphi}_{ji}(X_{1,j})\right)^2\middle|\tilde{\mathbf{X}}_n\right]\wedge\varepsilon\right]}{\varepsilon} \\
&\to_n 0,
\end{aligned}$$

using that $x \mapsto x \min \varepsilon$ is concave and the dominated convergence theorem. This proves that

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\hat{\varphi}_{ji}(X_{k,j})\right)^2 \xrightarrow{P}_n \mathbb{E}[X_{1,i}-\varphi_{i,j}(X_{1,j}))^2].$$

Similar arguments also show that the second term of Equation (C.10) also converges to zero in probability. Thus, we have shown that the second term of Equation (C.9) converges to zero in probability. Finally, the above arguments apply similarly to

the first term of Equation (C.9) by exchanging every φ_{ji} with $\tilde{\varphi}_{ji}$. We have shown that $\sup_{\tilde{\mathcal{G}}\in\mathcal{T}_p} |\hat{\ell}_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell^*_{\mathrm{G}}(\tilde{\mathcal{G}})| \xrightarrow{P}_n 0$, which concludes the proof.

□

Proof of Theorem 4.2: For simplicity of the proof, we assume that $\mathbb{E}[X] = 0$ such that the edge weight estimators simplify to

$$\hat{w}_{\mathrm{G}}(j \to i) := \hat{w}_{ji}(\mathbf{X}_n, \tilde{\mathbf{X}}_n) := \frac{1}{2}\log\left(\frac{\frac{1}{n}\sum_{k=1}^n \left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j})\right)^2}{\frac{1}{n}\sum_{k=1}^n X_{k,i}^2}\right).$$

Let $\ell := \ell_{\mathrm{G}}$ and $\hat{\ell} := \hat{\ell}_{\mathrm{G}}$ for notational simplicity. We know that for each SCM θ_n it holds that

$$\ell(\mathcal{G}) + q_n \leq \ell(\tilde{\mathcal{G}}),$$

for all $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$. Thus, it suffices to show that

$$\begin{aligned}
&P_{\theta_n}\left(\operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p} \sum_{(j\to i)\in\tilde{\mathcal{E}}} \hat{w}_{ji} = \mathcal{G}\right) \\
&= P_{\theta_n}\left(\left(|\hat{\ell}(\mathcal{G}) - \ell(\mathcal{G})| < \frac{q_n}{2}\right) \cap \bigcap_{\tilde{\mathcal{G}}\in\mathcal{T}_p\setminus\{\mathcal{G}\}} \left(\hat{\ell}(\tilde{\mathcal{G}}) \geq \ell(\tilde{\mathcal{G}}) - \frac{q_n}{2}\right)\right) \to_n 1.
\end{aligned}$$

For any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$ we have that

$$\hat{\ell}(\tilde{\mathcal{G}}) - \ell(\tilde{\mathcal{G}}) = \sum_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \hat{w}_{ji} - w_{ji} + \sum_{(j\to i)\notin\tilde{\mathcal{E}}\setminus\mathcal{E}} \hat{w}_{ji} - w_{ji},$$

where $\hat{w}_{ji}$ and w_{ji} denotes the estimated and population Gaussian weights for the edge $(j \to i)$. Hence, it suffices to show that

$$\begin{aligned}
&\forall (j \to i) \in \mathcal{E}, \forall \varepsilon > 0 : P_{\theta_n}(|\hat{w}_{ji} - w_{ji}| < q_n\varepsilon) \to_n 1, \\
&\forall (j \to i) \notin \mathcal{E}, \forall \varepsilon > 0 : P_{\theta_n}\left(\hat{w}_{ji} - w_{ji} \geq -q_n\varepsilon\right) \to_n 1.
\end{aligned}$$

To see this, note that in the affirmative, then

$$\begin{aligned}
P_{\theta_n}\left(|\hat{\ell}(\mathcal{G}) - \ell(\mathcal{G})| < \frac{q_n}{2}\right) &\geq P_{\theta_n}\left(\sum_{(j\to i)\in\mathcal{E}} |\hat{w}_{ji} - w_{ji}| < \frac{q_n}{2}\right) \\
&\geq P_{\theta_n}\left(\bigcap_{(j\to i)\in\mathcal{E}} \left(|\hat{w}_{ji} - w_{ji}| < \frac{q_n}{2(p-1)}\right)\right) \\
&\to_n 1,
\end{aligned}$$

and for any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$

$$\begin{aligned}
&P_{\theta_n}\left(\hat{\ell}(\tilde{\mathcal{G}}) - \ell(\tilde{\mathcal{G}}) \geq -\frac{q_n}{2}\right) \\
&= P_{\theta_n}\left(\sum_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \hat{w}_{ji} - w_{ji} + \sum_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \hat{w}^n_{ji} - w_{ji} \geq -\frac{q_n}{2}\right) \\
&\geq P_{\theta_n}\left(\bigcap_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \left(|\hat{w}_{ji} - w_{ji}| \leq \frac{q_n}{2(p-1)}\right)\right. \\
&\qquad\qquad \left.\bigcap_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \left(\hat{w}_{ji} - w_{ji} \geq -\frac{q_n}{2(p-1)}\right)\right) \\
&\to_n 1,
\end{aligned}$$

hence the probability of the intersections also converges to one.

The causal edges: Now fix $(j \to i) \in \mathcal{E}$, we want to show that for all $\varepsilon > 0$ it holds that

$$P_{\theta_n}(|\hat{w}_{ji} - w_{ji}| < q_n\varepsilon) \to_n 1.$$

First note that

$$\begin{aligned}
|\hat{w}_{ji} - w_{ji}| \leq \frac{1}{2}\left|\log\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\right) - \log\left(\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]\right)\right| \\
+ \frac{1}{2}\left|\log(\mathbb{E}[X_i^2]) - \log\left(\frac{1}{n}\sum_{k=1}^n X_{k,i}^2\right)\right|,
\end{aligned}$$

and note that it is clear that it suffices to show the wanted convergence in probability for each of the above terms. Furthermore, for a sequence of positive random variables (Z_n) and positive constant $c > 0$ then for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$P_\theta(q_n^{-1}|\log(Z_n) - \log(c)| > \varepsilon) \leq P_\theta(q_n^{-1}|Z_n - c| > \delta),$$

for sufficiently large n. To see this, note that if $q_n^{-1}(\log(Z_n) - \log(c)) > \varepsilon$, then $Z_n > \exp(\log(c) + q_n\varepsilon) = c\exp(q_n\varepsilon) \geq c(1 + q_n\varepsilon)$, so $q_n^{-1}(Z_n - c) > c\varepsilon$. On the other hand, if $q_n^{-1}(\log(Z_n) - \log(c)) < -\varepsilon$, then $Z_n < c\exp(-\varepsilon q_n) \leq c(1 - \varepsilon q_n + \varepsilon^2 q_n^2)$, so $q_n^{-1}(Z_n - c) < -c\varepsilon + c\varepsilon^2 q_n$. In summary, if $q_n^{-1}|\log(Z_n) - \log(c)| > \varepsilon$, then $q_n^{-1}|Z_n - c| > c\varepsilon - c\varepsilon^2 q_n > c\varepsilon(1 - M) =: \delta$ where $1 > M > \varepsilon q_n$ for sufficiently large n. We conclude that it suffices to show that for all $\varepsilon > 0$ it holds that

$$P_{\theta_n}\left(\left|\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2 - \mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]\right| \geq q_n\varepsilon\right) \to_n 0 \qquad \text{(C.12)}$$

and that

$$P_{\theta_n}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_{k,i}^2 - \mathbb{E}[X_i^2]\right| \geq q_n\varepsilon\right) \to_n 0, \tag{C.13}$$

Equation (C.13) is easily seen to be satisfied as the terms are mean zero i.i.d. such that

$$W_n := \frac{1}{n}\sum_{k=1}^{n} X_{k,i}^2 - \mathbb{E}[X_i^2],$$

where $\mathbb{E}_{\theta_n}[q_n^{-1}W_n] = 0$ and that $\mathbb{E}_{\theta_n}[q_n^{-2}W_n^2] = \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}[(X_i^2 - \mathbb{E}[X_i^2])^2]$, hence

$$\begin{aligned} P_{\theta_n}(q_n^{-1}W_n \geq \varepsilon) &\leq q_n^{-2}\frac{\mathbb{E}_{\theta_n}[W_n^2]}{\varepsilon^2} \\ &\leq \frac{q_n^{-2}}{n}\frac{\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}[(X_i^2 - \mathbb{E}[X_i^2])^2]}{\varepsilon^2} \\ &\to_n 0, \end{aligned}$$

for any $\varepsilon > 0$ as $\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}\|X\|_2^4 < \infty$ and $q_n^{-1} = o(\sqrt{n})$.

Now we show Equation (C.12). First, we simplify the notation by letting $Z_k := X_{k,i}$, $Y_k := X_{k,j}$ $f := \varphi_{ji}$ and $\hat{f} := \hat{\varphi}_{ji}$. Note that we have suppressed the dependence of $f = \varphi_{ji}$ on θ_n. We have that

$$\begin{aligned} \frac{1}{n}\sum_{k=1}^{n}\left(Z_k - \hat{f}(Y_k)\right)^2 &= \frac{1}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))^2 + \frac{1}{n}\sum_{k=1}^{n}(f(Y_k) - \hat{f}(Y_k))^2 \\ &\quad + \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)) \\ &=: T_{1,n} + T_{2,n} + T_{3,n}, \end{aligned}$$

and note that it suffices to show that for all $\varepsilon > 0$ it holds that

(a) $P_{\theta_n}\left(|T_{1,n} - \mathbb{E}[(Z_1 - f(Y_1))^2]| \geq q_n\varepsilon\right) \to_n 0$,

(b) $P_{\theta_n}\left(|T_{2,n}| \geq q_n\varepsilon\right) \to_n 0$,

(c) $P_{\theta_n}\left(|T_{3,n}| \geq q_n\varepsilon\right) \to_n 0$.

First we show (a). We note that each term in the sum of $T_{1,n} - \mathbb{E}[(Z_1 - f(Y_1))^2]$ is mean zero and i.i.d., i.e.,

$$q_n^{-1}\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]] = 0.$$

Furthermore,

$$\begin{aligned}
&\mathrm{Var}_{\theta_n}(q_n^{-1}(T_{1,n} - \mathbb{E}[(Z_1 - f(Y_1))^2])) \\
=&\mathrm{Var}_{\theta_n}\left(\frac{q_n^{-1}}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]\right) \\
=&\frac{q_n^{-2}}{n^2}\sum_{k=1}^{n}\mathrm{Var}_{\theta_n}\left((Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]\right) \\
\leq&\frac{q_n^{-2}}{n}\sup_{n\in\mathbb{N}}\mathrm{Var}_{\theta_n}\left((Z_1 - f(Y_1))^2\right) \\
\to_n&0,
\end{aligned}$$

since $q_n^{-1} = o(\sqrt{n})$ and that $\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}\|X\|_2^4 < \infty$. Hence,

$$\begin{aligned}
&P_{\theta_n}\left(|q_n^{-1}(T_{1,n} - \mathbb{E}[(Z_1 - f(Y_1))^2])| \geq \varepsilon\right) \\
&\leq \frac{\mathrm{Var}_{\theta_n}(q_n^{-1}(T_{,n} - \mathbb{E}[(Z_1 - f(Y_1))^2]))}{\varepsilon^2} \\
&\to_n 0.
\end{aligned}$$

by Markov's inequality, proving (a).

Now we show (b). To that end, simply note that the terms of $T_{2,n}$ is i.i.d. conditional on $\tilde{\mathbf{X}}_n$, hence fix $\varepsilon > 0$ and note that

$$\begin{aligned}
P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) &= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(q_n^{-1}T_{2,n} \geq \varepsilon|\tilde{\mathbf{X}}_n\right) \wedge 1\right] \\
&\leq \frac{\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-1}T_{2,n}|\tilde{\mathbf{X}}_n\right] \wedge 1\right]}{\varepsilon} \\
&= \frac{\mathbb{E}_{\theta_n}\left[q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right]}{\varepsilon},
\end{aligned}$$

where we used the conditional Markov's inequality. Now let $\delta > 0$ and define $A_{n,\delta} := (q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] > \delta)$ and note that by assumption there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Hence, for $n \geq N_\delta$ we have that

$$\begin{aligned}
&\mathbb{E}_{\theta_n}\left[q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&= \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \qquad \text{(C.14)} \\
&\quad + \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}^c}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\leq \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\quad + \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}^c}\delta\right] \\
&\leq \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}\right] + \delta \\
&= P_{\theta_n}(A_{n,\delta}) + \delta < 2\delta, \qquad \text{(C.15)}
\end{aligned}$$

hence $\limsup_{n\to\infty} P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) < 2\delta/\varepsilon$, i.e., $P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) \to 0$ as $\delta > 0$ was chosen arbitrarily, proving (b).

Now we prove (c). To this end, recall that

$$T_{3,n} := \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)),$$

is, conditional on $\tilde{\mathbf{X}}$, an i.i.d. sum with conditional mean zero

$$\begin{aligned}\mathbb{E}_{\theta_n}[T_{3,n}|\tilde{\mathbf{X}}_n] &= 2\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] \\ &= 2\mathbb{E}_{\theta_n}[(\mathbb{E}_{\theta_n}[Z_k|Y_k, \tilde{\mathbf{X}}_n] - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] \\ &= 2\mathbb{E}_{\theta_n}[(f(Y_k) - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] = 0,\end{aligned}$$

and conditional second moment given by

$$\begin{aligned}\mathbb{E}_{\theta_n}[T_{3,n}^2|\tilde{\mathbf{X}}_n] &= \frac{4}{n^2}\sum_{k=1}^{n}\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))^2(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[(Z_k - f(Y_k))^2(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[(Z_k - f(Y_k))^2|\tilde{\mathbf{X}}_n, Y_k\right](f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[\mathrm{Var}_{\theta_n}(Z_k|Y_k)(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &\leq \frac{C}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right]\end{aligned}$$

P_{θ_n}-almost surely. Hence, the conditional Markov's inequality yields that

$$\begin{aligned}P_{\theta_n}(q_n^{-1}T_{3,n} \geq \varepsilon) &= \mathbb{E}_{\theta_n}[P_{\theta_n}(q_n^{-1}T_{3,n} \geq \varepsilon|\tilde{\mathbf{X}}_n) \wedge 1] \\ &\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}T_{3,n}^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{C}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right].\end{aligned} \tag{C.16}$$

By conditional Jensen's inequality, we have that

$$\begin{aligned}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right]] &\leq 1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right]^2 \\ &\leq 1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right].\end{aligned}$$

Fix $\delta > 0$ and let $A_{n,\delta} := \left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] > \delta\right)$ and note that $P_{\theta_n}(A_{n,\delta}) \to_n 0$, hence there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Furthermore, as $q_n^{-1} = o(\sqrt{n})$ there exists an $N \in \mathbb{N}$ such that $q_n^{-2}/n < \delta$ for all

$n \geq N$. Similar to the arguments in Equation (C.15) we then have that

$$\begin{aligned}\frac{\varepsilon^2}{C} P_{\theta_n}(q_n^{-1} T_{3,n} \geq \varepsilon) &\leq \mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\left(1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4 | \tilde{\mathbf{X}}_n\right]\right) \wedge 1\right] \\ &\leq \frac{q_n^{-2}}{n} + \mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4 | \tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{q_n^{-2}}{n} + \mathbb{E}_{\theta_n}[1_{A_{n,\delta}}] + \mathbb{E}[1_{A_{n,\delta}^c}\delta] \\ &< \delta + P_{\theta_n}(A_{n,\delta}) + \delta < 3\delta,\end{aligned}$$

for any $n \geq N_\delta \vee N$, so $P_{\theta_n}(q_n^{-1} T_{3,n} \geq \varepsilon) \to_n 0$, proving (c).

The non-causal edges: Now fix $(j \to i) \notin \mathcal{E}$, we want to show, for any $\varepsilon > 0$ that

$$P_{\theta_n}(\hat{w}_{ji} - w_{ji} \geq -q_n \varepsilon) \to_n 1,$$

where

$$\begin{aligned}\hat{w}_{ji} - w_{ji} = \frac{1}{2}\Bigg(&\log\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\right) - \log\left(\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]\right) \\ &+ \log(\mathbb{E}[X_i^2]) - \log\left(\frac{1}{n}\sum_{k=1}^n X_{k,i}^2\right)\Bigg) =: \frac{1}{2}(D_{1,n} - D_{2,n}).\end{aligned}$$

We have that $P_{\theta_n}(\hat{w}_{ji} - w_{ji} \geq -q_n\varepsilon) \geq P_{\theta_n}\left((D_{1,n} \geq -q_n\varepsilon) \cap (|D_{2,n}| < q_n\varepsilon)\right)$, where the second event have already been show to have probability converging to one. Thus, it suffices to show that

$$P_{\theta_n}\left(D_{1,n} \geq -q_n\varepsilon\right) \to_n 1.$$

By similar arguments as above we have for any sequence of positive random variables $(K_n)_{n\geq 1}$ and a positive constant K that for all $\varepsilon > 0$ there exists an $\delta > 0$ such that $P_{\theta_n}\left(\log(K_n) - \log(K) < -q_n\varepsilon\right) \leq P_{\theta_n}(K_n - K < -q_n\delta)$, for sufficiently large $n \in \mathbb{N}$. To see this, note that if $\log(K_n) - \log(K) < -q_n\varepsilon$, then $K_n < K\exp(-\varepsilon q_n) \leq K(1 - \varepsilon q_n + \varepsilon^2 q_n^2)$, so $q_n^{-1}(K_n - K) < -K\varepsilon + K\varepsilon^2 q_n < -K\varepsilon(1 - M) =: -\delta$ where $1 > M > \varepsilon q_n$ for sufficiently large n as $q_n \downarrow 0$. Thus, it suffices to show that for any $\varepsilon > 0$ it holds that

$$P_{\theta_n}\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2 - \mathbb{E}_{\theta_n}[(X_i - \varphi_{ji}(X_j))^2] \geq -q_n^{-1}\varepsilon\right) \to_n 1.$$

Again, we simplify the notation $Z_k := X_{k,i}$, $Y_k := X_{k,j}$, $f = \varphi_{ji}$ and $\hat{f} := \hat{\varphi}_{ji}$. Now define the following terms

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(Z_k - \hat{f}(Y_k)\right)^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2] \\
&= \frac{1}{n}\sum_{k=1}^{n}\{(Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]\} \\
&\quad + \frac{1}{n}\sum_{k=1}^{n}\{(f(Y_k) - \hat{f}(Y_k))^2 - \delta^2_{n,\theta_n}\} \\
&\quad + \frac{2}{n}\sum_{k=1}^{n}\{(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)) + \delta^2_{n,\theta_n}/2\} \\
&=: \tilde{T}_{1,n} + \tilde{T}_{2,n} + \tilde{T}_{3,n},
\end{aligned}$$

where $\delta^2_{n,\theta_n} := \mathbb{E}_{\theta_n}[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n] = \mathbb{E}_{\theta_n}[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^2|\tilde{\mathbf{X}}_n]$ and note that it suffices to show that for all $\varepsilon > 0$ it holds that

(d) $P_{\theta_n}\left(|\tilde{T}_{1,n} - \mathbb{E}[Z_1 - f(Y_1)]| \geq q_n\varepsilon\right) \to_n 0$,

(e) $P_{\theta_n}\left(|\tilde{T}_{2,n}| \geq q_n\varepsilon\right) \to_n 0$,

(f) $P_{\theta_n}\left(\tilde{T}_{3,n} \geq -q_n\varepsilon\right) \to_n 1$.

Condition (d) holds by arguments similar to (a) for the causal edges.

Now we prove (e). To see this, note that it is, conditional on $\tilde{\mathbf{X}}_n$, a sum of mean zero i.i.d. terms, hence

$$\begin{aligned}
&\mathbb{E}_{\theta_n}\left(q_n^{-2}\tilde{T}^2_{2,n}\middle|\,\tilde{\mathbf{X}}_n\right) \\
&= \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[\{(f(Y_k) - \hat{f}(Y_k))^2 - \delta^2_{n,\theta_n}\}^2|\tilde{\mathbf{X}}_n\right] \\
&= \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4 + (\delta^2_{n,\theta_n})^2 - 2(f(Y_k) - \hat{f}(Y_k))^2\delta^2_{n,\theta_n}|\tilde{\mathbf{X}}_n\right] \\
&= \frac{q_n^{-2}}{n}\left(\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] - (\delta^2_{n,\theta_n})^2\right) \\
&\leq \frac{q_n^{-2}}{n}2\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right].
\end{aligned}$$

Here we used the Conditional Jensen's inequality, i.e., that we have

$$(\delta^2_{n,\theta_n})^2 \leq \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right].$$

Fix $\delta > 0$ and let $A_{n,\delta} := \left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] > \delta\right)$ and note that there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Similar to the previous

arguments we have for any $\varepsilon > 0$ and $n \geq N_\delta$ that

$$\begin{aligned}
P_{\theta_n}\left(\left|\tilde{T}_{2,n}\right| \geq q_n\varepsilon\right) &= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(\left|q_n^{-1}\tilde{T}_{2,n}\right| \geq \varepsilon \middle| \tilde{\mathbf{X}}_n\right) \wedge 1\right] \\
&\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}\tilde{T}_{2,n}^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\leq \frac{2}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\leq \frac{2}{\varepsilon^2}\left(\mathbb{E}_{\theta_n}[1_{A_{n,\delta}}] + \mathbb{E}_{\theta_n}[1_{A_{n,\delta}^c}\delta]\right) < \frac{4\delta}{\varepsilon^2},
\end{aligned}$$

by the conditional Markov's inequality. Since $\delta > 0$ was chosen arbitrarily, we conclude that (e) holds.

Finally we show (f). Recall that from the analysis of the causal edges, we defined

$$T_{3,n} := \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)).$$

Hence, we have that $\tilde{T}_{3,n} = T_{3,n} + \delta_{n,\theta_n}^2$. We realize that

$$\begin{aligned}
P_{\theta_n}(\tilde{T}_{3,n} < -q_n\varepsilon) &\leq P_{\theta_n}(T_{3,n} + \delta_{n,\theta_n}^2 \leq -q_n\varepsilon) \\
&= P_{\theta_n}\left(T_{3,n} \leq -\left(q_n\varepsilon + \delta_{n,\theta_n}^2\right)\right) \\
&\leq P_{\theta_n}\left(T_{3,n}^2 \geq \left(q_n\varepsilon + \delta_{n,\theta_n}^2\right)^2\right) \\
&\leq P_{\theta_n}\left(T_{3,n}^2 \geq (q_n\varepsilon)^2\right) \\
&= P_{\theta_n}\left(q_n^{-2}T_{3,n}^2 \geq \varepsilon^2\right) \\
&= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(q_n^{-2}T_{3,n}^2 \geq \varepsilon^2|\tilde{\mathbf{X}}_n\right) \wedge 1\right] \\
&\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}T_{3,n}^2\middle|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\to_n 0,
\end{aligned}$$

where we used the convergence shown in the proof of (c); see Equation (C.16). To see that the former arguments apply to non-causal edges, simply note that the former arguments did not use any conditions restricted to causal edges. This concludes the proof.

□

C.4.3. Proofs of Section 4.4

Lemma C.1. *Consider an i.i.d. sequence $(X_m)_{m\geq 1}$ of random variables with $X_m \in \mathbb{R}^d$ independent from a random infinite sequence $\tilde{\mathbf{X}} \in \prod_{i=1}^{\infty}\mathbb{R}^d$. Let $(\psi_n)_{n\geq 1}$ be a sequence of measurable functions with $\psi_n : \mathbb{R}^d \times (\prod_{i=1}^{\infty}\mathbb{R}^d) \to \mathbb{R}^q$ for all $n \geq 1$ satisfying the following conditions:*

(a) $\mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}] = 0$ *almost surely,*

(b) $\sum_{m=1}^{n} \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) \xrightarrow{P}_n \Sigma$*,*

(c) $\sum_{m=1}^{n} \mathbb{E}[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}|\tilde{\mathbf{X}}] \xrightarrow{P}_n 0$ *for some* $\varepsilon > 0$*.*

It holds that

$$\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

Proof of Lemma C.1: Let the random sequences be defined on a common probability space $(\Omega, \mathcal{F}, P)$ and define

$$\begin{aligned} A_{nm} &:= \mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}], \\ B_n &:= \Sigma - \sum_{m=1}^{n} \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}), \\ C_n &:= \sum_{m=1}^{n} \mathbb{E}[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}|\tilde{\mathbf{X}}]. \end{aligned}$$

By assumption we have that $P(\cap_{n,m}(A_{nm} = 0)) = 1, B_n \xrightarrow{P} 0$ and $C_n \xrightarrow{P} 0$ as $n \to \infty$. First, note that for any subsequence $(n_k)_{k\geq 1}$ of the positive integers, there exists a further subsequence $(n_{k_l})_{l\in\mathbb{N}}$ such that

$$(\lim_{l\to\infty} B_{n_{k_l}} = 0) := \{\omega \in \Omega : \lim_{l\to\infty} B_{n_{k_l}}(\omega) = 0\}, \quad \text{with} \quad P(\lim_{l\to\infty} B_{n_{k_l}} = 0) = 1.$$

and

$$(\lim_{l\to\infty} C_{n_{k_l}} = 0) := \{\omega \in \Omega : \lim_{l\to\infty} C_{n_{k_l}}(\omega) = 0\}, \quad \text{with} \quad P(\lim_{l\to\infty} C_{n_{k_l}} = 0) = 1.$$

Thus, define

$$G := (\cap_{n,m}(A_{nm} = 0) \cap (\lim_{l\to\infty} B_{n_{k_l}} = 0) \cap (\lim_{l\to\infty} C_{n_{k_l}} = 0) \subseteq \Omega, \quad \text{with} \quad P(G) = 1.$$

Now fix $\tilde{x} \in \tilde{\mathbf{X}}(G) = \{\tilde{\mathbf{X}}(\omega) \in \prod_{j=1}^{\infty} \mathbb{R}^p : \omega \in G\}$ and note that for $l \geq 1$ we have that

$$\begin{aligned} &\forall 1 \leq m \leq n_{k_l} : \mathbb{E}[\psi_{n_{k_l}}(X_m, \tilde{x})] = 0, \\ &\sum_{m=1}^{n_{k_l}} \mathrm{Var}(\psi_{n_{k_l}}(X_m, \tilde{x})) \to \Sigma, \\ &\sum_{m=1}^{n_{k_l}} \mathbb{E}[\|\psi_{n_{k_l}}(X_m, \tilde{x})\|_2^{2+\varepsilon}] \to 0. \end{aligned}$$

Furthermore as

$$\psi_{n_{k_l}}(X_1, \tilde{x}) \perp\!\!\!\perp \cdots \perp\!\!\!\perp \psi_{n_{k_l}}(X_{n_{k_l}}, \tilde{x}),$$

for any $l \geq N_{\tilde{x}}$ we have by Lyapunov's central limit theorem for triangular arrays (see, e.g., Van der Vaart, 2000, Proposition 2.27, and recall that Lyapunov's condition implies the Lindeberg–Feller condition) that

$$\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{x}) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0, \Sigma).$$

The above convergence in distribution is equivalent to the following: for any continuous bounded function $g : \mathbb{R}^{p^2} \to \mathbb{R}$ it holds that

$$\lim_{v \to \infty} \mathbb{E}\left[g\left(\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{x}) \right) \right] = \mathbb{E}\left[g(Z) \right].$$

Fix a continuous and bounded g and note that the above convergence holds for all $\tilde{x} \in \tilde{\mathbf{X}}(G)$ with $P(G) = 1$. Thus, it must hold that

$$\mathbb{E}\left[g\left(\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{\mathbf{X}}) \right) \Big| \tilde{\mathbf{X}} \right] \xrightarrow{a.s.} \mathbb{E}\left[g(Z) \right].$$

Finally, as $(n_{k_l})_{l \geq 1}$ was chosen as a further subsequence of an arbitrary subsequence of positive integers, we have that

$$\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{x}) \right) \Big| \tilde{\mathbf{X}} \right] \xrightarrow{P} \mathbb{E}\left[g(Z) \right],$$

and since g is bounded the dominated convergence theorem yields that

$$\begin{aligned} &\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \right) \right] \\ =&\mathbb{E}\left[\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \right) \Big| \tilde{\mathbf{X}} \right] \right] \to \mathbb{E}\left[g(Z) \right]. \end{aligned}$$

As g was chosen arbitrarily, the above convergence holds for any continuous bounded g. We conclude that

$$\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

proving the theorem. □

Lemma C.2 (Shah and Peters, 2020, Lemma 19)**.** *Let $\mathcal{P}$ be a family of distributions for a random variable $\zeta \in \mathbb{R}$ and suppose $\zeta_1, \zeta_2, \ldots$ are i.i.d. copies of ζ. For each $n \in \mathbb{N}$ let $S_n = n^{-1} \sum_{i=1}^{n} \zeta_i$. Suppose that for all $P \in \mathcal{P}$ we have $\mathbb{E}_P(\zeta) = 0$ and $\mathbb{E}_P\left(|\zeta|^{1+\eta}\right) < c$ for some $\eta, c > 0$. We have that for all $\varepsilon > 0$,*

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}_n} \mathbb{P}_P\left(|S_n| > \varepsilon \right) = 0.$$

Lemma C.3. *Let U be a random element and let $(Z_n)_{n\geq 1}$ be an i.i.d. sequence of random variables such that $U \perp\!\!\!\perp (Z_n)_{n\geq 1}$ and let $((W_{nm})_{m\geq n})_{n\geq 1}$ be a triangular array of random variables and $(g_n)_{n\geq 1}$ be measurable mappings with the following properties:*

1. *for each n, m, $W_{nm} := g_n(Z_m, U)$,*
2. *for some $\eta > 0, \mathbb{E}\left(|W_{n1}|^{1+\eta} \mid U\right) = O_p(1)$*

Then writing $\bar{W}_n := \sum_{i=1}^n W_{ni}/n$, we have

$$\left|\bar{W}_n - \mathbb{E}\left(W_n \mid U\right)\right| \xrightarrow{p} 0.$$

Proof of Lemma C.3: Denote

$$j_n(Z_m, U) := g_n(Z_m, U) - \mathbb{E}[g_n(Z_1, U)|U],$$

for any $m \leq n$ and $n \geq 1$. Let $\delta > 0$ be given. Pick $M > 0$ and $N \in \mathbb{N}$ such that the events

$$\Omega_n := \left\{\mathbb{E}\left[|g_n(Z_{n1}, U)|^{1+\eta} \mid U\right] \leq M\right\},$$

satisfy $\mathbb{P}\left(\Omega_n^c\right) < \delta$ for $n \geq N$. Notice that

$$U(\Omega_n) = \left\{\tilde{u}_n : \mathbb{E}\left[|g_n(Z_1, \tilde{u}_n)|^{1+\eta}\right] \leq M\right\},$$

and that

$$\begin{aligned}\mathbb{P}\left(\left|\bar{W}_n - \mathbb{E}\left(W_n \mid U\right)\right| > \varepsilon\right) &= P\left(\left|\frac{1}{n}\sum_{m=1}^n j_n(Z_m, U)\right| > \varepsilon\right) \\ &< \mathbb{E}\left[P\left(\left|\frac{1}{n}\sum_{m=1}^n j_n(Z_m, U)\right| > \varepsilon \mid U\right) 1_{\Omega_n}\right] + \delta.\end{aligned}$$

By the dominated convergence theorem, the first term on the RHS will converge to 0 if

$$\begin{aligned}&\sup_{\omega\in\Omega_n} P\left(\left|\frac{1}{n}\sum_{m=1}^n j_n(Z_m, U)\right| > \varepsilon \mid U\right)(\omega) \\ &= \sup_{\tilde{u}_n \in U(\Omega_n)} P\left(\left|\frac{1}{n}\sum_{m=1}^n j_n(Z_m, \tilde{u}_n)\right| > \varepsilon\right) \to 0,\end{aligned}$$

which implies the desired conclusion as $\delta > 0$ was chosen arbitrarily. Now note that for any $\tilde{u}_n \in U(\Omega_n)$ it holds that

$$\mathbb{E}[|j_n(Z_i, \tilde{u}_n)|^{1+\eta}] \leq M, \quad \text{and} \quad \mathbb{E}[j_n(Z_i, \tilde{u}_n)] = 0,$$

for all $i \in \mathbb{N}$. Now let $(Y_i)_{i\geq 1}$ be a sequence of i.i.d. random variables such that for each background probability measure $P' \in \mathcal{P}_n$ is holds that $Y_1 \overset{\mathcal{D}}{=} j_n(Z_1, \tilde{u}_n)$ for some $\tilde{u}_n \in U(\Omega_n)$ such that $E|Y_1|^{1+\eta} \leq M$ and $E[Y_1] = 0$. Thus,

$$\begin{aligned}\sup_{\tilde{u}_n \in U(\Omega_n)} P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, \tilde{u})\right| > \varepsilon\right) &= \sup_{P' \in \mathcal{P}_n} P'\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > \varepsilon\right) \\ &\leq \sup_{P' \in \cup\mathcal{P}_n} P'\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > \varepsilon\right) \\ &\to_n 0,\end{aligned}$$

by the weak uniform law of large numbers, Lemma C.2. □

Lemma C.4 (Asymptotic normality of edge weight components)**.** *Let $\hat{\varphi}^n_{ji}$ denote the estimated conditional mean function φ_{ji} based on the sample $\tilde{\mathbf{X}}_n$. For any $j \neq i$ and $m \leq n$, define*

$$\begin{aligned}&\hat{R}_{nm,ji} := \{X_{m,i} - \hat{\varphi}^n_{ji}(X_{m,j})\}, && \hat{\mu}_{n,ji} := \frac{1}{n}\sum_{m=1}^{n} \hat{R}^2_{nm,ji}, \\ &R_{m,ji} := \{X_{m,i} - \varphi_{ji}(X_{m,j})\}, && \mu_{ji} := \mathbb{E}[R^2_{1,ji}], \\ &\hat{V}_{m,i} := \left(X_{m,i} - \frac{1}{n}\sum_{k=1}^{n} X_{k,i}\right)^2, && \hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^{n} \hat{V}_{m,i} \\ &\nu_i := \mathrm{Var}(X_{1,i}), && \delta^2_{n,ji} := \mathbb{E}[(\hat{\varphi}_{ji}(X_{1,j}) - \varphi_{ji}(X_{1,j}))^2 | \tilde{\mathbf{X}}_n].\end{aligned}$$

and

$$\widehat{\Sigma}_n := \begin{bmatrix} \widehat{\Sigma}_{n,R} & \widehat{\Sigma}_{n,RV} \\ \widehat{\Sigma}^\intercal_{n,RV} & \widehat{\Sigma}_{n,V} \end{bmatrix} := \frac{1}{n}\sum_{m=1}^{n} \begin{bmatrix} \hat{R}^2_{nm}(\hat{R}^2_{nm})^\intercal - \hat{\mu}_n\hat{\mu}_n^\intercal & \hat{R}^2_{nm}\hat{V}_m^\intercal - \hat{\mu}_n\hat{\nu}_n^\intercal \\ \hat{V}_m(\hat{R}^2_{nm})^\intercal - \hat{\nu}_n\hat{\mu}_n^\intercal & \hat{V}_m\hat{V}_m^\intercal - \hat{\nu}_n\hat{\nu}_n^\intercal \end{bmatrix},$$

denote an $p^2 \times p^2$ matrix empirical covariance matrix, where the squared vectors denote that each entry is squared. Suppose there exists $\xi > 0$ such that for all $j \neq i$, the following three conditions hold:

(i) $\mathbb{E}\|X\|^{4+\xi} < \infty$.

(ii) $\mathbb{E}[|\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j)|^{4+\xi} | \tilde{\mathbf{X}}_n] = O_p(1)$, *as* $n \to \infty$.

(iii) $\mathrm{Var}\left(\begin{bmatrix} \hat{R}^2_{n1} - \delta^2_n - \mu \\ \hat{V}_1 - \nu \end{bmatrix} \middle| \tilde{\mathbf{X}}_n\right) \xrightarrow{P}_n \Sigma$, *where* Σ *is constant.*

Then we have that $\widehat{\Sigma}_n \xrightarrow{P} \Sigma \in \mathbb{R}^{p^2 \times p^2}$ *and that*

$$\frac{1}{\sqrt{n}}\sum_{m=1}^{n} \begin{bmatrix} \hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu \end{bmatrix} = \sqrt{n}\begin{bmatrix} \hat{\mu}_n - \delta^2_n - \mu \\ \hat{\nu}_n - \nu \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \tag{C.17}$$

Proof of Lemma C.4: We prove the lemma under the assumption that $\mathbb{E}[X] = 0$ for which we can simplify the variance estimator by $\hat{V}_{m,i} := X_{m,i}^2$ and $\hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^{n} \hat{V}_{m,i}$ for all $1 \leq i \leq p$. The proof only gets more notionally cumbersome without this assumption. I.e., it should follow in all generality by applying expansion techniques and Slutsky's theorem similar to the standard arguments showing asymptotic normality of the regular sample variance.

Note that when conditioning $\hat{\varphi}_{ji}^n$ on $\tilde{\mathbf{X}}$ it is equivalent to conditioning on $\tilde{\mathbf{X}}_n$ by the i.i.d. structure of $\tilde{\mathbf{X}}$ and that $\hat{\varphi}_{ji}^n$ only depends on $\tilde{\mathbf{X}}_n$, the first n coordinates of $\tilde{\mathbf{X}}$.

First, we define for all $j \neq i$, $m \leq n$ and $n \in \mathbb{N}$ the following conditional expectation regression error $\hat{\delta}_{nm,ji} := \{\varphi_{ji}(X_{m,j}) - \hat{\varphi}_{ji}^n(X_{m,j})\}$. Furthermore, for each $n \in \mathbb{N}$ and $m \leq n$ define

$$\Psi_n(X_m, \tilde{\mathbf{X}}) := \begin{bmatrix} \hat{R}_{nm}^2 - \delta_n^2 - \mu \\ \hat{V}_m - \nu \end{bmatrix} \in \mathbb{R}^{p^2},$$

where only $\tilde{\mathbf{X}}_n$ (the first n coordinates of $\tilde{\mathbf{X}}$) is used, and

$$\psi_n(X_m, \tilde{\mathbf{X}}) := \frac{1}{\sqrt{n}} \Psi_n(X_m, \tilde{\mathbf{X}}).$$

Note that the desired conclusion of Equation (C.17) follows by verifying condition (a), (b) and (c) of Lemma C.1. First, we show (a), the conditional mean zero condition. To that end, note that for any $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, p\} \setminus \{i\}$ it holds that

$$\begin{aligned} \hat{R}_{nm,ji}^2 &= (X_{m,i} - \varphi_{ji}(X_{m,j}) + \varphi_{ji}(X_{m,j}) - \hat{\varphi}_{ji}^n(X_{m,j}))^2 \\ &= (R_{m,ji} + \hat{\delta}_{nm,ji})^2 \\ &= R_{m,ji}^2 + \hat{\delta}_{nm,ji}^2 + 2R_{m,ji}\hat{\delta}_{nm,ji}. \end{aligned}$$

Hence, we have that

$$\hat{R}_{nm,ji}^2 - \mu_{ji} - \delta_{n,ji}^2 = (R_{m,ji}^2 - \mu_{ji}) + (\hat{\delta}_{nm,ji}^2 - \delta_{n,ji}^2) + 2R_{m,ji}\hat{\delta}_{nm,ji}. \tag{C.18}$$

The terms of Equation (C.18) are mean zero conditionally on $\tilde{\mathbf{X}}$, since $\mathbb{E}[R_{m,ji}^2|\tilde{\mathbf{X}}] = \mathbb{E}[R_{m,ji}^2] = \mu_{ji}$, $\mathbb{E}[\hat{\delta}_{nm,ji}^2|\tilde{\mathbf{X}}] = \delta_{n,ji}^2$ and

$$\begin{aligned} \mathbb{E}[R_{m,ji}\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] &= \mathbb{E}[\mathbb{E}[R_{m,ji}\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}, X_{m,j}]|\tilde{\mathbf{X}}] \\ &= \mathbb{E}[\mathbb{E}[X_{m,i} - \varphi_{ji}(X_{m,j})|\tilde{\mathbf{X}}, X_{m,j}]\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] \\ &= \mathbb{E}[(\mathbb{E}[X_{m,i}|X_{m,j}] - \varphi_{ji}(X_{m,j}))\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] \\ &= 0, \end{aligned}$$

as $\varphi_{ji}(X_{m,j}) = \mathbb{E}[X_{m,i}|X_{m,j}]$ almost surely. Furthermore,

$$\mathbb{E}[X_{m,i}^2 - \mathrm{Var}(X_i)|\tilde{\mathbf{X}}] = \mathbb{E}[X_{m,i}^2] - \mathrm{Var}(X_i) = 0.$$

We conclude that

$$\mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\begin{bmatrix}\hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu\end{bmatrix}\middle|\tilde{\mathbf{X}}\right] = 0,$$

almost surely. With respect to (b), convergence of the sum of variances, we by assumption have that

$$\Sigma_n := \begin{bmatrix}\Sigma_{n,R} & \Sigma_{n,RV} \\ \Sigma_{n,RV}^\intercal & \Sigma_{n,V}\end{bmatrix} := \mathrm{Var}\left(\Psi_n(X_1, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}\right) \xrightarrow{P}_n \Sigma, \tag{C.19}$$

where as Σ is a positive semi-definite matrix. Furthermore, we have that $(X_m)_{m\geq 1}$ is an i.i.d. sequence independent of $\tilde{\mathbf{X}}$. Therefore,

$$\begin{aligned}\sum_{m=1}^n \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) &= \sum_{m=1}^n \frac{1}{n}\mathrm{Var}(\Psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) \\ &= \sum_{m=1}^n \frac{1}{n}\Sigma_n \\ &= \Sigma_n \\ &\xrightarrow{P} \Sigma.\end{aligned}$$

Finally, we show that condition (c), a conditional Lindeberg-Feller condition, is fulfilled. To this end, note that with $\varepsilon = \xi/2 > 0$ we have that

$$\begin{aligned}&\mathbb{E}\left[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\begin{bmatrix}\hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu\end{bmatrix}\right\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \qquad \text{(C.20)} \\ &= \frac{1}{n^{\frac{2+\varepsilon}{2}}}\mathbb{E}\left[\left\|\begin{bmatrix}\hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu\end{bmatrix}\right\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \\ &\leq \frac{1}{n^{\frac{2+\varepsilon}{2}}}2^{(\frac{2+\varepsilon}{2}-1)}\Bigg(\sum_{i\neq j}\mathbb{E}\left[|\hat{R}^2_{nm,ji} - \mu_{ji} - \delta^2_{n,ji}|^{2+\varepsilon}|\tilde{\mathbf{X}}\right] \\ &\qquad + \sum_{i=1}^p \mathbb{E}|X^2_{m,i} - \mathrm{Var}(X_i)|^{2+\varepsilon}\Bigg), \qquad \text{(C.21)}\end{aligned}$$

by the cr and quadratic form inequalities. We now realize that the latter factor of Equation (C.21) is stochastically bounded. To see this, note that for any $j \neq i$ it holds that

$$\begin{aligned}\mathbb{E}\left[|\hat{R}^2_{nm,ji} - \mu_{ji} - \delta^2_{n,ji}|^{2+\varepsilon}|\tilde{\mathbf{X}}\right] \leq 2^{1+\varepsilon}(&\mathbb{E}[|\hat{R}_{nm,ji}|^{4+2\varepsilon}|\tilde{\mathbf{X}}] + \mu_{ji}^{2+\varepsilon} \\ &+ \mathbb{E}[|\delta^2_{n,ji}(\tilde{\mathbf{X}})|^{2+\varepsilon}|\tilde{\mathbf{X}}]). \qquad \text{(C.22)}\end{aligned}$$

The first term of the upper bound in Equation (C.22) is $O_p(1)$,

$$\begin{aligned}
&\mathbb{E}[|\hat{R}_{nm,ji}|^{4+2\varepsilon}|\tilde{\mathbf{X}}] \\
&= \mathbb{E}[|X_{m,i} - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}] \\
&\leq 2^{3+2\varepsilon}(\mathbb{E}|X_{m,i} - \varphi_{ji}(X_{m,j})|^{4+2\varepsilon} + \mathbb{E}[|\varphi_{ji}(X_{m,i}) - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}]) \\
&= 2^{3+2\varepsilon}(\mathbb{E}[|R_{m,ji}|^{4+\xi}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]) = O_p(1),
\end{aligned}$$

as $\mathbb{E}\|X\|_2^{4+\xi} < \infty$ and $\mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] = O_p(1)$. That is, $R_{m,ji} = \{X_{m,i} - \mathbb{E}[X_{m,i}|X_{m,j}]\}$ of which both terms are in $\mathcal{L}^{4+\xi}(P)$ if $X_{m,i} \in \mathcal{L}^{4+\xi}(P)$ which is guaranteed as $\mathbb{E}\|X\|_2^{4+\xi} < \infty$. For the thid term in the upper bound of Equation (C.22), we note that by the conditional Jensen's inequality, we have that

$$\begin{aligned}
\mathbb{E}[|\delta^2_{n,ji}|^{2+\varepsilon}|\tilde{\mathbf{X}}] &\leq \mathbb{E}[|\varphi_{ji}(X_{m,i}) - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}] \\
&= \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] \\
&= O_p(1),
\end{aligned}$$

by assumption. Therefore, since $n^{\frac{2+\varepsilon}{2}} = n^{1+\varepsilon/2} > n$ we have that

$$\sum_{m=1}^{n} \mathbb{E}\left[\|\Psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \leq \frac{n}{n^{\frac{2+\varepsilon}{2}}} O_p(1) = n^{-\varepsilon/2} O_p(1) \xrightarrow{P} 0,$$

proving the conditional Lindeberg-Feller condition. By Lemma C.1 it holds that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

Now it only remains to prove that

$$\|\widehat{\Sigma}_n - \Sigma_n\| \xrightarrow{P} 0,$$

or, equivalently, that each entry converges to zero in probability. For example, for the entries of the first block matrix with $j \neq i$ and $l \neq r$ we prove that

$$|\widehat{\Sigma}_{n,R,ji,lr} - \Sigma_{n,R,ji,lr}| \xrightarrow{P} 0.$$

Now note that the observable estimated covariance matrix entry is given by

$$\widehat{\Sigma}_{n,R,ji,lr} = \frac{1}{n} \sum_{m=1}^{n} \hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr} - \hat{\mu}_{n,ji} \hat{\mu}_{n,lr},$$

while the unobservable conditional covariance matrix is given by

$$\begin{aligned}
\Sigma_{n,R,ji,lr} &= \mathbb{E}[(\hat{R}^2_{nm,ji} - \mu_{ji} - \delta^2_{n,ji})(\hat{R}^2_{nm,lr} - \mu_{lr} - \delta^2_{n,lr})|\tilde{\mathbf{X}}] \\
&= \mathbb{E}[\hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - (\mu_{n,ji} + \delta^2_{n,ji})(\mu_{n,lr} + \delta^2_{n,lr}) \\
&= \mathbb{E}[\hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}],
\end{aligned}$$

Note that the second term of the feasible covariance matrix estimator expands to

$$\begin{aligned}
\hat{\mu}_{n,ji}\hat{\mu}_{n,lr} &= \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji}\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr}\right) \\
&= \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji} - \mathbb{E}[\hat{R}^2_{nm,ji}]\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,lr}]\right) \\
&\quad - \mathbb{E}[\hat{R}^2_{nm,ji}]\mathbb{E}[\hat{R}^2_{nm,lr}] \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji}\mathbb{E}[\hat{R}^2_{nm,lr}] \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr}\mathbb{E}[\hat{R}^2_{nm,ji}],
\end{aligned}$$

Thus

$$\begin{aligned}
&|\Sigma_{n,R,ji,lr} - \widehat{\Sigma}_{n,R,ji,lr}| \\
=&\Bigg|\frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]) \\
&\quad - \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji} - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]\right) \\
&\quad - \frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,ji}\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]) \\
&\quad - \frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,lr}\mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}])\Bigg|. \qquad \text{(C.23)}
\end{aligned}$$

Each of these terms tends to zero in probability by Lemma C.3. For example, for the first term of Equation (C.23) it suffices to show that

$$\mathbb{E}\left[|\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|^{1+\varepsilon}|\tilde{\mathbf{X}}\right] = O_p(1),$$

for some $\varepsilon > 0$. Fix $\varepsilon = \xi/4$ and note that

$$\begin{aligned}
\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr} &= (X_{m,i} - \hat{\varphi}^n_{ji}(\tilde{\mathbf{X}})(X_{m,j}))^2(X_{m,r} - \hat{\varphi}^n_{lr}(\tilde{\mathbf{X}})(X_{m,l}))^2 \\
&\leq 4(R^2_{m,ji} + \hat{\delta}^2_{nm,ji})(R^2_{m,lr} + \hat{\delta}^2_{nm,lr}).
\end{aligned}$$

Thus, by the cr-inequality and the conditional Cauchy-Schwarz inequality we have

that

$$\begin{aligned}
&\mathbb{E}[|\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|^{1+\varepsilon}|\tilde{\mathbf{X}}] \\
\leq& 4\mathbb{E}[(R^2_{m,ji}+\hat{\delta}^2_{nm,ji})^{1+\varepsilon}(R^2_{m,lr}+\hat{\delta}^2_{nm,lr})^{1+\varepsilon}|\tilde{\mathbf{X}}] \\
\leq& 4 2^{2\varepsilon}\mathbb{E}[(|R_{m,ji}|^{2+2\varepsilon}+|\hat{\delta}_{nm,ji}|^{2+2\varepsilon})(|R_{m,lr}|^{2+2\varepsilon}+|\hat{\delta}_{nm,lr}|^{2+2\varepsilon})|\tilde{\mathbf{X}}] \\
\leq& \mathbb{E}[|R_{m,ji}|^{2+2\varepsilon}|R_{m,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] + \mathbb{E}[|R_{m,ji}|^{2+2\varepsilon}|\hat{\delta}_{nm,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] \\
&\quad + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{2+2\varepsilon}|R_{m,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{2+2\varepsilon}|\hat{\delta}_{nm,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] \\
\leq& \mathbb{E}[|R_{m,ji}|^{4+\xi}]\mathbb{E}[|R_{m,lr}|^{4+\xi}] + \mathbb{E}[|R_{m,ji}|^{4+\xi}]\mathbb{E}[|\hat{\delta}_{nm,lr}|^{4+\xi}|\tilde{\mathbf{X}}] \\
&\quad + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]\mathbb{E}[|R_{m,lr}|^{4+\xi}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]\mathbb{E}[|\hat{\delta}_{nm,lr}|^{4+\xi}|\tilde{\mathbf{X}}] \\
=& O_p(1),
\end{aligned}$$

as $\mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] = O_p(1)$ for all $j \neq i$ by assumption and $\mathbb{E}[|R_{m,ji}|^{4+\xi}] < \infty$ since $\mathbb{E}\|X\|_2^{4+\xi} < \infty$.

Similar arguments show convergence in probability of the entries in the other block submatrices of $\widehat{\Sigma}_n$ less Σ_n, yielding the desired conclusion.

□

Proof of Theorem 4.3: We prove the theorem under the simplifying assumption that $\mathbb{E}[X] = 0$ for which we can simplify the variance estimator by $\hat{V}_{m,i} := X^2_{m,i}$ and $\hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^n \hat{V}_{m,i}$ for all $1 \leq i \leq p$.

First, note (using the notation introduced in Lemma C.4) that $\hat{M}_1 = \{\hat{R}^2_{n1,ji}\}_{j\neq i}$ for which the conditional mean given $\tilde{\mathbf{X}}_n$ is given by

$$\mathbb{E}[\hat{M}_1|\tilde{\mathbf{X}}_n] = \mathbb{E}[\{\hat{R}^2_{n1,ji}\}_{j\neq i}|\tilde{\mathbf{X}}_n] = \mu + \delta^2_n,$$

see Equation (C.18). Similarly we have that $\mathbb{E}[\hat{V}_1|\tilde{\mathbf{X}}_n] = \mathbb{E}[\hat{V}_1] = \nu$. Subtracting a constant (conditional on $\tilde{\mathbf{X}}_n$) does not change the conditional variance, hence

$$\mathrm{Var}\left(\begin{bmatrix}\hat{R}^2_{n1} - \delta^2_n - \mu \\ \hat{V}_1 - \nu\end{bmatrix}\middle|\tilde{\mathbf{X}}_n\right) = \mathrm{Var}\left((\hat{M}_1^\intercal, \hat{V}_1^\intercal)^\intercal\middle|\tilde{\mathbf{X}}_n\right) \xrightarrow{P}_n \Sigma,$$

where Σ is constant and positive semi-definite. As such, we satisfy the conditions of Lemma C.4 which yields that

$$\frac{1}{\sqrt{n}}\sum_{m=1}^n \begin{bmatrix}\hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu\end{bmatrix} = \sqrt{n}\begin{bmatrix}\hat{\mu} - \delta^2_n - \mu \\ \hat{\nu} - \nu\end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \tag{C.24}$$

and that

$$\widehat{\Sigma} = \begin{bmatrix}\widehat{\Sigma}_M & \widehat{\Sigma}_{MV} \\ \widehat{\Sigma}_{MV}^\intercal & \widehat{\Sigma}_V\end{bmatrix} \xrightarrow{P} \Sigma =: \begin{bmatrix}\Sigma_M & \Sigma_{MV} \\ \Sigma_{MV}^\intercal & \Sigma_V\end{bmatrix} \in \mathbb{R}^{p^2\times p^2}.$$

For any $j \neq i$ we denote

$$\hat{w}_{ji} = \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right), \quad \tilde{w}_{ji} = \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji} - \delta^2_{n,ji}}{\hat{\nu}_i}\right), \quad w_{ji} = \frac{1}{2}\log\left(\frac{\mu_{ji}}{\nu_i}\right),$$

where the latter is simply a shorthand notation for the Gaussian edge weight $w_{\mathrm{G}}(j \to i)$. Fix $\alpha \in (0,1)$. First, consider $(j \to i) \in \mathcal{E}$ and note that

$$\begin{aligned}\sqrt{n}\left(\begin{bmatrix}\hat{\mu}_{ji} - \mu_{ji} \\ \hat{\nu}_i - \nu_i\end{bmatrix} - \begin{bmatrix}\hat{\mu}_{ji} - \delta^2_{n,ji} - \mu_{ji} \\ \hat{\nu}_i - \nu_i\end{bmatrix}\right) &= \sqrt{n}\begin{bmatrix}\delta^2_{n,ji} \\ 0\end{bmatrix} \\ &= \sqrt{n}\begin{bmatrix}\mathbb{E}[\hat{\delta}^2_{nm,ji}|\tilde{\mathbf{X}}_n] \\ 0\end{bmatrix} \\ &\xrightarrow{P}_n 0.\end{aligned}$$

Hence, the delta method yields that

$$\begin{aligned}&\sqrt{n}(\log(\hat{\mu}_{ji}) - \log(\mu_{ji}) - \log(\hat{\nu}_i) + \log(\nu_i)) \\ =&\sqrt{n}\left(\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right) - \log\left(\frac{\mu_{ji}}{\nu_i}\right)\right) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \tilde{\sigma}^2_{ji}),\end{aligned}$$

where

$$\hat{\sigma}^2_{ji} := \frac{\widehat{\Sigma}_{M,ji}}{\hat{\mu}^2_{ji}} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}^2_i} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{\hat{\mu}_{ji}\hat{\nu}_i} \xrightarrow{P} \sigma^2_{ji} := \frac{\Sigma_{M,ji}}{\mu^2_{ji}} + \frac{\Sigma_{V,i}}{\nu^2_i} - 2\frac{\Sigma_{MV,ji,i}}{\mu_{ji}\nu_i} \geq 0.$$

Here $\widehat{\Sigma}_{M,ji}$ and $\widehat{\Sigma}_{V,i}$ and their limits use a shorthand notation that denote the corresponding diagonal element, e.g., $\widehat{\Sigma}_{M,ji} := \widehat{\Sigma}_{M,ji,ji}$.

An asymptotically valid marginal confidence interval for w_{ji} with level α is, by virtue of the above convergence in distribution, given by

$$\hat{w}_{ji} \pm \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}},$$

where $q(1-\frac{\alpha}{2})$ is the $1-\alpha/2$ quantile of the standard normal distribution. That is,

$$P\left(\hat{w}_{ji} - \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}} \leq w_{ji} \leq \hat{w}_{ji} + \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}}\right) \xrightarrow{P} 1-\alpha.$$

On the other hand, for any $(j \to i) \notin \mathcal{E}$ we have, by similar arguments, except that no assumption guarantees that $\sqrt{n}\delta^2_{n,ji}$ vanishes, that

$$P\left(\tilde{w}_{ji} - \tilde{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}} \leq w_{ji} \leq \tilde{w}_{ji} + \tilde{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}}\right) \xrightarrow{P} 1-\alpha,$$

where

$$\begin{aligned}\tilde{\sigma}_{ji}^2 &:= \frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)\hat{\nu}_i} \\ &\xrightarrow{P} \sigma_{ji}^2 := \frac{\Sigma_{M,ji}}{\mu_{ji}^2} + \frac{\Sigma_{V,i}}{\nu_i^2} - 2\frac{\Sigma_{MV,ji,i}}{\mu_{ji}\nu_i} \geq 0.\end{aligned}$$

Note that $\tilde{\sigma}_{ji}^2$ is not observable since $\delta_{n,ji}^2$ is not observable. Thus, we have the following Bonferroni corrected simultaneous confidence interval for the Gaussian edge weights

$$\begin{aligned}\liminf_{n\to\infty} P\Bigg(&\bigcap_{(j\to i)\in\mathcal{E}} \left(w_{ji} \in \left[\hat{w}_{ji} \pm \hat{\sigma}_{ji}\frac{q\left(1-\frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}}\right]\right) \\ &\bigcap_{j\to i\notin\mathcal{E}} \left(w_{ji} \in \left[\tilde{w}_{ji} \pm \tilde{\sigma}_{ji}\frac{q\left(1-\frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}}\right]\right)\Bigg) \geq 1-\alpha.\end{aligned}$$

The above confidence region has the correct asymptotic level, but it is infeasible as $\tilde{w}_{ji}$, $\tilde{\sigma}_{ji}$ and $\mathcal{E}$ are not observable. Let, for all $j \neq i$, $\hat{l}_{\alpha,ji}$, $\hat{u}_{\alpha,ji}$ and $\tilde{l}_{\alpha,ji}$, $\tilde{u}_{\alpha,ji}$ denote the lower and upper bounds of the confidence intervals using $\hat{w}_{ji}$, $\hat{\sigma}_{ji}$ and $\tilde{w}_{ji}$, $\tilde{\sigma}_{ji}$, respectively. To see this, note that

$$\begin{aligned}C(\hat{l}_\alpha, \tilde{l}_\alpha, \hat{u}_\alpha, \tilde{u}_\alpha) := \Bigg\{ &\underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji} : \forall (j\to i)\in\mathcal{E}, w'_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}], \\ &\forall (j\to i)\notin\mathcal{E}, w'_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}]\Bigg\},\end{aligned}$$

is an unobservable confidence region for the causal graph. That is,

$$\begin{aligned}&\liminf_{n\to\infty} P(\mathcal{G} \in C(\hat{l}_\alpha, \tilde{l}_\alpha, \hat{u}_\alpha, \tilde{u}_\alpha)) \\ \geq &\liminf_{n\to\infty} P\left(\bigcap_{(j\to i)\in\mathcal{E}} (w_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]) \bigcap_{(j\to i)\notin\mathcal{E}} (w_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}])\right) \\ \geq &1-\alpha.\end{aligned}$$

On the other hand, our proposed confidence region has the form

$$\hat{C} := \hat{C}(\hat{l}_\alpha, \hat{u}_\alpha) := \left\{ \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji} : \forall j \neq i, w'_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]\right\},$$

which corresponds to the biased but feasible confidence region

$$\prod_{j\neq i}\left[\hat{w}_{ji} \pm \hat{\sigma}_{ji}\frac{q\left(1-\frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}}\right] =: \prod_{j\neq i}[\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}].$$

for the Gaussian edge weights. The biased confidence region $\prod_{j\neq i}[\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]$ does not necessarily contain the population Gaussian edge weights with a probability of at least $1-\alpha$ in the large sample limit. However, it can be used to construct a conservative confidence region for the causal graph. To see this, note that it is clear that by further penalizing the wrong edge weights, the causal graph will still yields the minimum edge weight spanning directed tree. Hence,

$$\begin{aligned}
&\liminf_{n\to\infty} P(\mathcal{G} \in \hat{C}(\hat{l}_\alpha, \hat{u}_\alpha)) \\
\geq &\liminf_{n\to\infty} P\Bigg(\bigcap_{(j\to i)\in\mathcal{E}} (w_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]) \bigcap_{(j\to i)\notin\mathcal{E}} (w_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}]) \\
&\qquad \bigcap_{(j\to i)\notin\mathcal{E}} (\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}) \Bigg) \\
\geq &1-\alpha,
\end{aligned}$$

as $P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to 1$ for all $(j \to i) \notin \mathcal{E}$ by Lemma C.5. □

Lemma C.5. *Suppose that the assumptions of Lemma C.4 hold. It holds that*

$$\forall (j \to i) \notin \mathcal{E}, \forall \alpha \in (0,1) : P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to_n 1.$$

Proof of Lemma C.5: Fix any $(j \to i) \notin \mathcal{E}$ and $\alpha \in (0,1)$ and note that we want to show that

$$\begin{aligned}
&\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji} \\
\iff &\tilde{w}_{ji} + c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}} \leq \hat{w}_{ji} + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} \\
\iff &0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta^2_{n,ji}\right) - c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}},
\end{aligned}$$

holds with probability converging to one, where c is a strictly positive constant. It suffices to show that an even smaller quantity is positive with increasing probability. That is,

$$0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta^2_{n,ji}\right) - c\frac{\tilde{\sigma}^*_{ji}}{\sqrt{n}},$$

with increasing probability, where

$$\tilde{\sigma}^*_{ji} := \sqrt{\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta^2_{n,ji})^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + 2\frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - \delta^2_{n,ji})\hat{\nu}_i}} \geq \tilde{\sigma}_{ji}.$$

with $\tilde{\sigma}^*_{ji} > 0$ (with increasing probability). Let $d_n(t) : [0,\infty) \to \mathbb{R}$ denote the random function given by

$$\begin{aligned}
d_n(t) := &\log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - t\right) \\
&- \frac{c}{\sqrt{n}}\sqrt{\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + 2\frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)\hat{\nu}_i}}.
\end{aligned}$$

It holds that $d_n(0) = 0$ surely, so by the mean value theorem, the desired conclusion holds if it with increasing probability (as n tends to infinity) holds that $d_n'(t) \geq 0$ for all $t \in [0, \delta_{n,ji}^2]$.

Now fix $\eta > 0$ and choose $M_\eta, \varepsilon_1, \ldots, \varepsilon_5 > 0$ such that the lower bounds in the following inequalities is positive

$$\begin{aligned}
\Omega_n(1) &:= (\hat{\mu}_{ji} \leq M_\eta), \\
\Omega_n(2) &:= (\Sigma_{M,ji} - \varepsilon_1 \leq \widehat{\Sigma}_{M,ji} \leq \Sigma_{M,ji} + \varepsilon_1), \\
\Omega_n(3) &:= (\Sigma_{V,i} - \varepsilon_2 \leq \widehat{\Sigma}_{V,i} \leq \Sigma_{V,i} + \varepsilon_2), \\
\Omega_n(4) &:= (0 \leq |\widehat{\Sigma}_{MV,ji,i}| \leq |\Sigma_{MV,ji,i}| + \varepsilon_3), \\
\Omega_n(5) &:= (\mu_{ji} - \varepsilon_4 \leq \hat{\mu}_{ji} - \delta_{n,ji}^2 \leq \mu_{ji} + \varepsilon_4), \\
\Omega_n(6) &:= (\nu_i - \varepsilon_5 \leq \hat{\nu}_i \leq \nu_i + \varepsilon_5),
\end{aligned}$$

and that $\liminf_{n\to\infty} P(\Omega_n(1)) > 1 - \eta$. This is possible as $\hat{\mu}_{n,ji} - \delta_{n,ji}^2 \xrightarrow{P} \mu_{ji} > 0$ and that

$$\begin{aligned}
\hat{\delta}_{n,ji}^2 = E[|\hat{\delta}_{nm,ji}|^2 | \tilde{\mathbf{X}}] &= E[|\hat{\delta}_{nm,ji}|^{\frac{4+\xi}{2+\xi/2}} | \tilde{\mathbf{X}}] \\
&\leq E[|\hat{\delta}_{nm,ji}|^{4+\xi} | \tilde{\mathbf{X}}]^{\frac{1}{2+\xi/2}} = O_p(1),
\end{aligned}$$

by the conditional Jensen's inequality and concavity of $[0, \infty) \ni x \mapsto x^{\frac{1}{2+\xi/2}}$, which implies that $\hat{\mu}_{ji} = \hat{\mu}_{ji} - \hat{\delta}_{n,ji}^2 + \hat{\delta}_{n,ji}^2 = o_p(1) + O_p(1) = O_p(1)$. Furthermore, note that as

$$\begin{aligned}
&\widehat{\Sigma}_{M,ji} \xrightarrow{P} \Sigma_{M,ji} > 0, \quad \widehat{\Sigma}_{V,i} \xrightarrow{P} \Sigma_{V,i} > 0, \\
&|\widehat{\Sigma}_{MV,ji,i}| \xrightarrow{P} |\Sigma_{MV,ji,i}| \geq 0, \quad \hat{\nu}_i \xrightarrow{P} \nu_i > 0,
\end{aligned}$$

it holds that

$$\begin{aligned}
\limsup_{n\to\infty} P\left(\bigcup_{1\leq k\leq 6} \Omega_n(k)^c \right) &\leq \sum_{1\leq k\leq 6} \limsup_{n\to\infty} P(\Omega_n(k)^c) \\
&= \limsup_{n\to\infty} P(\Omega_n(1)^c) \leq \eta.
\end{aligned}$$

Here we used that the diagonal elements of the limit covariance matrix is strictly positive. The fact that $\mu_{ji}, \nu_i > 0$ follows from the fact that $X_i - \mathbb{E}[X_i | X_j]$ is assumed to have density (w.r.t. Lebesgue measure) and that the variables are non-degenerate $\nu_i = \mathrm{Var}(X_i) > 0$. Thus, we have that

$$\liminf_{n\to\infty} P\left(\bigcap_{1\leq k\leq 6} \Omega_n(k) \right) \geq 1 - \eta.$$

Now consider a fixed $\omega \in \bigcap_{1\leq k\leq 6} \Omega_n(k)$ and note that with $g_n : [0, \delta_{n,ji}^2] \to \mathbb{R}$ given by $g_n(t) = \hat{\mu}_{ji} - t$ we have that g_n is decreasing and that

$$g_n([0, \delta_{n,ji}^2]) \subseteq [\mu_{ji} - \varepsilon_4, \hat{\mu}_{ji}] \subseteq (0, M_\delta]$$

We have that for any $t \in [0, \delta_{n,ji}^2]$ that

$$d_n'(t) = \frac{1}{\hat{\mu}_{ji} - t} - \frac{2c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + \frac{2|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)\hat{\nu}_i} \right)^{-1/2} \times \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)^2 \hat{\nu}_i} \right),$$

hence,

$$\begin{aligned} d_n'(t) &= \frac{1}{\hat{\mu}_{ji} - t} - \frac{2c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{g_n(t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + \frac{2|\widehat{\Sigma}_{MV,ji,i}|}{g_n(t)\hat{\nu}_i} \right)^{-1/2} \\ &\quad \times \left(\frac{\widehat{\Sigma}_{M,ji}}{g_n(t)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{g_n(t)^2 \hat{\nu}_i} \right) \\ &\geq \frac{1}{\hat{\mu}_{ji}} - \frac{2c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{\hat{\mu}_{ji}^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} \right)^{-1/2} \\ &\quad \times \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)^2 \hat{\nu}_i} \right) \\ &\geq \frac{1}{M_\eta} - \frac{2c}{\sqrt{n}} \left(\frac{\Sigma_{M,ji} - \varepsilon_1}{M_\eta^2} + \frac{\Sigma_{V,i} - \varepsilon_2}{(\nu_i + \varepsilon_5)^2} \right)^{-1/2} \\ &\quad \times \left(\frac{\Sigma_{M,ji} + \varepsilon_1}{(\mu_{ji} - \varepsilon_4)^3} + \frac{|\Sigma_{MV,ji,i}| + \varepsilon_3}{(\mu_{ji} - \varepsilon_4)^2 (\nu_i - \varepsilon_5)} \right) \\ &=: \frac{1}{M_\eta} - \frac{C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5}}{\sqrt{n}} \\ &\geq 0, \end{aligned}$$

for $n \geq (C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5} / M_\eta)^2$. We conclude that

$$\begin{aligned} P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) &= P\left(0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta_{n,ji}^2\right) - c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}}\right) \\ &\geq P\left(\forall t \in [0, \delta_{n,ji}^2] : d_n'(t) \geq 0\right) \\ &\geq P\left(\bigcap_{1 \leq k \leq 6} \Omega_n(k)\right), \end{aligned}$$

for $n \geq (C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5} / M_\eta)^2$. Hence,

$$\liminf_{n \to \infty} P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \geq \liminf_{n \to \infty} P\left(\bigcap_{1 \leq k \leq 6} \Omega_n(k)\right) \geq 1 - \eta,$$

and as $\eta > 0$ was chosen arbitrarily, we finally have the desired conclusion

$$P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to 1.$$

□

Proof of Theorem 4.4: Consider a collection of arbitrary and possibly data-dependent substructures $\mathcal{R}_1, \mathcal{R}_2, ...$ and level $\alpha \in (0,1)$. First, we note that the score associated with two sets of edge weights w_1 and w_2 is weakly monotone, that is, $s(w_1) \leq s(w_2)$ if w_1 and w_2 satisfy the component-wise partial ordering $w_1 \leq w_2$. Furthermore, the restricted score function $w \mapsto s_{\mathcal{T}(\mathcal{R})}(w)$ is also weakly monotone for any set of restrictions $\mathcal{R}$.

Let $k \in \mathbb{N}$ and suppose that the null hypothesis

$$\mathcal{H}_0(\mathcal{R}_k) : \mathcal{E}_{\mathcal{R}_k} \setminus \mathcal{E} = \emptyset,\ \mathcal{E} \setminus \mathcal{E}^{\text{miss}}_{\mathcal{R}_k} = \emptyset,\ r_k = \text{rt}(\mathcal{G}),$$

corresponding to the restriction $\mathcal{R}_k = (\mathcal{E}_{\mathcal{R}_k}, \mathcal{E}^{\text{miss}}_{\mathcal{R}_k}, r_k)$ is true.

It is clear that, if there is a graph in $\hat{C} := \hat{C}(\hat{l}_\alpha, \hat{u}_\alpha)$ satisfying the restrictions imposed by the substructure $\mathcal{R}_k$, then there exist $\hat{l}_\alpha \leq w' \leq \hat{u}_\alpha$ such that $s(w')$ attains its minimum value in a graph satisfying $\mathcal{R}_k$. Now note that further penalizing (or removing) edges that are not present in the minimum edge weight directed tree does not affect the score of the minimum edge weigh directed tree. Hence, it holds that

$$s_{\mathcal{T}(\mathcal{R}_k)}(w') = s(w').$$

Monotonicity of $s_{\mathcal{T}(\mathcal{R}_k)}$ and s in the edge weights imply that

$$s_{\mathcal{T}(\mathcal{R}_k)}(\hat{l}_\alpha) \leq s_{\mathcal{T}(\mathcal{R}_k)}(w') = s(w') \leq s(\hat{u}_\alpha).$$

Hence, $s_{\mathcal{T}(\mathcal{R}_k)}(\hat{l}_\alpha) > s(\hat{u}_\alpha)$ entails that no graph in $\hat{C}$ satisfies the restrictions of $\mathcal{R}_k$. This is a slightly conservative criterion as $s_{\mathcal{T}(\mathcal{R}_k)}(\hat{l}_\alpha) \leq s(\hat{u}_\alpha)$ does not necessarily guarantee that a graph in $\hat{C}$ satisfies the restrictions of $\mathcal{R}_k$.

Therefore, if $\psi_{\mathcal{R}_k} = 1$, then we know that there is no graph in $\hat{C}$ satisfying the restrictions of $\mathcal{R}_j$. As the causal graph $\mathcal{G}$ satisfies the restriction $\mathcal{R}_k$ we conclude that $\mathcal{G}$ is not contained in $\hat{C}$. Thus for any true $\mathcal{R}_k$ we have that

$$(\psi_{\mathcal{R}_k} = 1) \subseteq (\mathcal{G} \notin \hat{C}).$$

Since this holds for any true $\mathcal{R}_k$, the conclusion follows by noting that

$$\limsup_{n\to\infty} P\left(\bigcup_{j:\mathcal{H}_0(\mathcal{R}_j) \text{ is true}} \{\psi_{\mathcal{R}_j} = 1\} \right) \leq \limsup_{n\to\infty} P(\mathcal{G} \notin \hat{C}) \leq \alpha,$$

where we used Theorem 4.3.

□

C.4.4. Proofs of Section 4.5

C.4.4.1. Proofs of first results in Section 4.5

Proof of Lemma 4.3: As conditioning reduces entropy we always have that

$$\begin{aligned}\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) = h(X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) &= h(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) \\ &\leq h(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]) \\ &= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i).\end{aligned}$$

Furthermore, note that when conditioning we throw out dependence information captured through the mutual information $I(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]; X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})$, which is zero if and only if $X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}] \perp\!\!\!\perp X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}$. This is especially the case for the true graph, i.e., $X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\mathcal{G}}(i)}] \perp\!\!\!\perp X_{\mathrm{pa}^{\mathcal{G}}(i)}$, implying that $\ell_{\mathrm{CE}}(\mathcal{G}, i) = \ell_{\mathrm{E}}(\mathcal{G}, i)$. Consequently, we have that the local conditional entropy score gap lower bounds the local entropy score gap,

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{CE}}(\mathcal{G}, i) \leq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i).$$

Furthermore, from the arguments in the proof of Lemma 4.2 we have that

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) &= \inf_{\tilde{N}_i \sim P_{\tilde{N}_i} \in \mathcal{P}} h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\ &\leq \inf_{\tilde{N}_i \sim P_{\tilde{N}_i} \in \mathcal{P}_G} h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\ &= \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) + \log(\sqrt{2\pi e}).\end{aligned}$$

If X is generated by a Gaussian noise model, i.e., with generating SCM $\theta = (\mathcal{G}, (f_i), P_N)$ with $P_N \in \mathcal{P}_{\mathrm{G}}^p$, then $\ell_{\mathrm{E}}(\mathcal{G}, i) = h(X_i - \mathbb{E}[X_i|X_{\mathrm{pa}^{\mathcal{G}}(i)}]) = h(N_i) = \log(\sqrt{2\pi e}\sigma_i) = \log(\sqrt{2\pi e}) + \frac{1}{2}\log(\mathbb{E}[N_i^2]) = \log(\sqrt{2\pi e}) + l_{\mathrm{G}}(\mathcal{G}, i)$, in which case the local entropy score gap lower bounds the local Gaussian score gap

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i) \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{G}}(\mathcal{G}, i).$$

□

Proof of Lemma 4.4: Note that $\mathbb{E}[Y|X] = \mathbb{E}[f(X)+N_Y|X] = f(X)+\mathbb{E}[N_Y|X] = f(X)+\mathbb{E}[N_Y]$, since $N_Y \perp\!\!\!\perp N_X = X$. Hence, the score difference can be written as

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, X) - \ell_{\mathrm{E}}(\mathcal{G}, X) + \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, Y) - \ell_{\mathrm{E}}(\mathcal{G}, Y) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(Y - E(Y|X)) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(N_Y + \mathbb{E}[N_Y]) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(N_Y),\end{aligned}$$

as the differential entropy is translation invariant. Now note that as $N_Y \perp\!\!\!\perp N_X$ it holds that $N_Y \perp\!\!\!\perp f(X)$, so conditioning on $f(X)$ yields that

$$\begin{aligned} h(Y) &= h(Y|f(X)) + I(Y; f(X)) \\ &= h(f(X) + N_Y|f(X)) + I(Y; f(X)) \\ &= h(N_Y) + I(Y; f(X)). \end{aligned}$$

Similarly, conditioning on X yields that

$$\begin{aligned} h(Y) &= h(Y|X) + I(Y; X) \\ &= h(N_Y) + I(Y; X), \end{aligned}$$

which proves that

$$I(Y; f(X)) = I(Y; X).$$

This equality is normally derived by restricting f to be bijective, but here it holds regardless by the structural assignment form, as Y is only dependent on X through $f(X)$. Furthermore, we have that

$$\begin{aligned} h(X - \mathbb{E}[X|Y]) &= I(X - \mathbb{E}[X|Y]; Y) + h(X - \mathbb{E}[X|Y]|Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) + h(X|Y). \end{aligned}$$

Hence,

$$\begin{aligned} h(X - \mathbb{E}[X|Y]) - h(X) &= I(X - \mathbb{E}[X|Y]; Y) + h(X|Y) - h(X) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X). \end{aligned}$$

Thus

$$\begin{aligned} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= h(X - \mathbb{E}[X|Y]) - h(X) + h(Y) - h(N_Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X) + h(N_Y) + I(Y; f(X)) - h(N_Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X) + I(Y; f(X)) \\ &= I(X - \mathbb{E}[X|Y]; Y), \end{aligned}$$

proving the claim. □

Proof of Proposition 4.2: As the conditional mean $\mathbb{E}[X|Y]$ vanishes, we have that

$$\begin{aligned} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= I(X - \mathbb{E}(X|Y); Y) \\ &= I(X; Y) \\ &= I(Y; X) \\ &= I(Y; f(X)), \end{aligned}$$

where the last equality was derived in the proof of Lemma 4.4. Alternatively, if we start with the entropy score gap definition we get by directed calculation that

$$\begin{aligned}
&h(N_X - \mathbb{E}[N_X|f(N_X)+N_Y]) - h(N_X) + h(f(N_X)+N_Y) - h(N_Y)\\
=&h(N_X) - h(N_X) + h(f(N_X)+N_Y) - h(N_Y)\\
=&h(f(N_X)+N_Y) - h(N_Y)\\
=&h(f(N_X)+N_Y|N_X) + I(f(N_X)+N_Y;N_X) - h(N_Y)\\
=&h(N_Y|N_X) + I(f(N_X)+N_Y;N_X) - h(N_Y)\\
=&h(N_Y) + I(f(N_X)+N_Y;N_X) - h(N_Y)\\
=&I(f(N_X)+N_Y;N_X)\\
=&I(Y;X) = I(Y,f(X)).
\end{aligned}$$

Now let $f(X)^{\mathrm{G}}$ and N_Y^{G} be independent normal distributed random variables with the same mean and variance as $f(X)$ and N_Y. That is,

$$f(X)^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[f(X)], \mathrm{Var}(f(X))), \quad \text{and} \quad N_Y^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[N_Y], \mathrm{Var}(N_Y)),$$

with $N_Y^{\mathrm{G}} \perp\!\!\!\perp f(X)^{\mathrm{G}}$ such that $f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[f(X)] + \mathbb{E}[N_Y], \mathrm{Var}(f(X)) + \mathrm{Var}(N_Y))$.

(a) If $D_{\mathrm{KL}}(f(X)\|f(X)^{\mathrm{G}}) \leq D_{\mathrm{KL}}(N_Y\|N_Y^{\mathrm{G}})$ then by Lemma C.1 of Silva (2009) we have, since $X \perp\!\!\!\perp N_Y$, that

$$I(Y;f(X)) = I(f(X)+N_Y;f(X)) \geq I(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}; f(X)^{\mathrm{G}}),$$

Note, we have equality if and only if $f(X)$ and N_Y are jointly Gaussian. Furthermore,

$$\begin{aligned}
&I(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}; f(X)^{\mathrm{G}})\\
&= h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}) - h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}|f(X)^{\mathrm{G}})\\
&= h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}) - h(N_Y^{\mathrm{G}})\\
&= \log(\sqrt{2\pi(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y))}) - \log(\sqrt{2\pi\mathrm{Var}(N_Y)})\\
&= \frac{1}{2}\log\left(\frac{\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)}{\mathrm{Var}(N_Y)}\right)\\
&= \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).
\end{aligned}$$

(b) If $f(X)+N_Y$ is log-concave distributed, then by Theorem 3 of Marsiglietti and Kostina (2018) we have that

$$\begin{aligned}
h(f(X)+N_Y) &\geq \frac{1}{2}\log\left(4\mathrm{Var}(f(X)+N_Y)\right)\\
&= \frac{1}{2}\log\left(4(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)\right).
\end{aligned}$$

Furthermore, it is well known that for fixed variance, the normal distribution maximizes entropy, hence

$$h(N_Y) \leq h(N_Y^{\mathrm{G}}) = \frac{1}{2}\log\left(2\pi \mathrm{Var}(N_Y)\right).$$

Therefore, we get that

$$\begin{aligned} I(Y; f(X)) &= I(f(X) + N_Y; f(X)) \\ &= h(f(X) + N_Y) - h(f(X) + N_Y | f(X)) \\ &= h(f(X) + N_Y) - h(N_Y) \\ &\geq \frac{1}{2}\log\left(4(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)\right) - \frac{1}{2}\log\left(2\pi e \mathrm{Var}(N_Y)\right) \\ &= \frac{1}{2}\log\left(\frac{2}{\pi e} + \frac{2}{\pi e}\frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right), \end{aligned}$$

which yields a strictly positive lower bound if and only if

$$\frac{2}{\pi e} + \frac{2}{\pi e}\frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)} > 1 \iff \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)} > \frac{\pi e}{2} - 1 \approx 3.27.$$

□

Lemma C.6. *Two different but Markov equivalent trees $\tilde{\mathcal{G}}$ and $\hat{\mathcal{G}}$ share the exact same edges except for a single reversed directed path between the two root nodes of the graphs,*

$$\begin{aligned} \hat{\mathcal{G}}: &\ c_1 \to c_2 \to \cdots \to c_{r-1} \to c_r, \\ \tilde{\mathcal{G}}: &\ c_r \to c_{r-1} \to \cdots \to c_2 \to c_1, \end{aligned}$$

with $c_1 = \mathrm{rt}(\hat{\mathcal{G}})$ and $c_r = \mathrm{rt}(\tilde{\mathcal{G}})$.

Proof of Lemma C.6: First, note that there always exists a unique directed path in $\hat{\mathcal{G}}$ from $\mathrm{rt}(\hat{\mathcal{G}})$ to $\mathrm{rt}(\tilde{\mathcal{G}})$

$$\hat{\mathcal{G}} : \mathrm{rt}(\hat{\mathcal{G}}) = c_1 \to \cdots \to c_{r-1} \to c_r = \mathrm{rt}(\tilde{\mathcal{G}}).$$

Since $\tilde{\mathcal{G}}$ and $\hat{\mathcal{G}}$ are Markov equivalent, they share the same skeleton, so in $\tilde{\mathcal{G}}$ the above path must be reversed. That is, there exists a unique directed path in $\tilde{\mathcal{G}}$ from $\mathrm{rt}(\tilde{\mathcal{G}})$ to $\mathrm{rt}(\hat{\mathcal{G}})$ given by

$$\tilde{\mathcal{G}} : \mathrm{rt}(\tilde{\mathcal{G}}) = c_r \to c_{r-1} \to \cdots \to c_1 = \mathrm{rt}(\hat{\mathcal{G}}),$$

If $r = p$ we are done, so assume $r < p$. As $\hat{\mathcal{G}}$ is a directed tree there must exists a node z_2 which is not a part of the above path but is a child of a node in the path. That is, there exists a node $z_1 \in \{c_1, \ldots, c_r\}$ such that $\hat{\mathcal{G}}$ contains the edge

$$\hat{\mathcal{G}} : z_1 \to z_2.$$

Furthermore, by equality of skeleton, this edge must also be present in $\tilde{\mathcal{G}}$,

$$\tilde{\mathcal{G}} : z_1 - z_2.$$

Assume for contradiction that $z_2 \to z_1$ in $\tilde{\mathcal{G}}$. As such, it must hold that $z_1 = c_r = \mathrm{rt}(\tilde{\mathcal{G}})$ for otherwise if $z_1 \in \{c_1, \ldots, c_{r-1}\}$ then z_1 would have two parents in $\tilde{\mathcal{G}}$, a contradiction since $\tilde{\mathcal{G}}$ is a directed tree. However, if $z_1 = c_r = \mathrm{rt}(\tilde{\mathcal{G}})$ then there is an incoming edge into the root node, a contradiction. We conclude that the directed edge $z_1 \to z_2$ also is present in $\tilde{\mathcal{G}}$.

Any paths further out on this branch will coincide in both graphs for otherwise there exists nodes with two parents. These arguments show that any paths branching out from the main reversed path will coincide in both $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$. Thus, the two graphs coincide up to a directed path between root nodes that is reversed. □

Proof of Proposition 4.3: By Lemma C.6 there exists a path reversal

$$\begin{array}{lccccccccc} \mathcal{G}: & \mathrm{rt}(\mathcal{G}) = c_1 & \to & c_2 & \to & \cdots & \to & c_{r-1} & \to & c_r = \mathrm{rt}(\tilde{\mathcal{G}}), \\ \tilde{\mathcal{G}}: & \mathrm{rt}(\tilde{\mathcal{G}}) = c_r & \to & c_{r-1} & \to & \cdots & \to & c_2 & \to & c_1 = \mathrm{rt}(\mathcal{G}), \end{array}$$

while all other edges in $\mathcal{G} = (V, \mathcal{E})$ and $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}})$ coincide. The entropy score gap is therefore only concerning the root nodes and the reversed edges in the above path. That is

$$\begin{aligned} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= h(X_{c_1}) - h(X_{c_r}) + \sum_{(j,i)\in\mathcal{E}} h(X_i - \mathbb{E}[X_i|X_j]) \\ &\quad - \sum_{(j,i)\in\tilde{\mathcal{E}}} h(X_i - \mathbb{E}[X_i|X_j]) \\ &= h(X_{c_1}) - h(X_{c_r}) + \sum_{i=1}^{r-1} h(X_{c_i} - \mathbb{E}[X_{c_i}|X_{c_{c+1}}]) \\ &\quad - \sum_{i=2}^{r} h(X_{c_i} - \mathbb{E}[X_{c_i}|X_{c_{i-1}}]). \end{aligned}$$

For easy of notation let $X_1 = X_{c_1}, \ldots, X_r = X_{c_r}$ and note that by Lemma 4.4 it holds that

$$\begin{aligned} h(X_{i+1}) + h(X_i - \mathbb{E}[X_i|X_{i+1}]) &= I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) \\ &\quad + h(X_i) + h(X_{i+1} - \mathbb{E}[X_{i+1}|X_i]). \end{aligned}$$

Thus,

$$
\begin{aligned}
&\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) + \sum_{i=1}^{r-2} h(X_{i+1}) \\
&= \left(\sum_{i=1}^{r-1} h(X_i - \mathbb{E}[X_i|X_{i+1}])\right) + h(X_r) + \sum_{i=1}^{r-2} h(X_{i+1}) \\
&= \sum_{i=1}^{r-1} h(X_i - \mathbb{E}[X_i|X_{i+1}]) + h(X_{i+1}) \\
&\geq \sum_{i=1}^{r-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) + \sum_{i=1}^{r-1} h(X_{i+1} - \mathbb{E}[X_{i+1}|X_i]) \\
&\quad + h(X_i) \\
&= \sum_{i=1}^{r-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) + \sum_{i=2}^{r} h(X_i - \mathbb{E}[X_i|X_{i-1}]) \\
&\quad + h(X_{i-1}) \\
&= \sum_{i=1}^{r-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) + \left(\sum_{i=2}^{r} h(X_i - \mathbb{E}[X_i|X_{i-1}])\right) \\
&\quad + h(X_1) + \sum_{i=3}^{r} h(X_{i-1}) \\
&= \sum_{i=1}^{r-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) + \left(\sum_{i=2}^{r} h(X_i - \mathbb{E}[X_i|X_{i-1}])\right) \\
&\quad + h(X_1) + \sum_{i=1}^{r-2} h(X_{i+1}) \\
&= \sum_{i=1}^{r-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) + \ell_{\mathrm{E}}(\mathcal{G}) + \sum_{i=1}^{r-2} h(X_{i+1}),
\end{aligned}
$$

proving that

$$
\begin{aligned}
\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq \sum_{i=1}^{p-1} I(X_i - \mathbb{E}[X_i|X_{i+1}]; X_{i+1}) \\
&= \sum_{i=1}^{r-1} \Delta\ell_{\mathrm{E}}(c_i \overleftarrow{\longrightarrow} c_{i+1}) \\
&\geq \min_{1\leq i\leq r-1} \Delta\ell_{\mathrm{E}}(c_i \overleftarrow{\longrightarrow} c_{i+1}).
\end{aligned}
$$

□

C.4.4.2. Proof of Theorem 4.5

We first describe the graphs that result from the reduction technique described in 4.5.3. To do so, define

$$\mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) := \{L \in V_R : \mathrm{ch}^{\mathcal{G}_R}(L) = \emptyset \wedge (\mathrm{pa}^{\mathcal{G}_R}(L) \neq \mathrm{pa}^{\mathcal{G}_R}(L) \vee \mathrm{ch}^{\mathcal{G}_R}(L) \neq \emptyset)\},$$

containing the sink nodes in $\mathcal{G}_R$ (these are either not sink nodes in $\tilde{\mathcal{G}}_R$ or are sink nodes in $\tilde{\mathcal{G}}_R$ with different parents: $\mathrm{pa}^{\mathcal{G}_R}(L) \neq \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L)$). Now fix any $L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \subseteq V_R$ and note that its only parent in $\mathcal{G}_R$, $\mathrm{pa}^{\mathcal{G}_R}(L)$, is either also a parent of L, a child of L or not adjacent to L, in $\tilde{\mathcal{G}}_R$. That is, one and only one of the following sets is non-empty

$$
\begin{aligned}
Z(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L), && \text{('staying parents')} \\
Y(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L), && \text{('parents to children')} \\
W(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L) \cup \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L)\}) && \text{('removing parents').}
\end{aligned}
$$

We define the $\tilde{\mathcal{G}}_R$ parent and children of L that are not adjacent to L in $\mathcal{G}_R$ as

$$
\begin{aligned}
D(L) &:= \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\mathcal{G}_R}(L) \cup \mathrm{pa}^{\mathcal{G}_R}(L)\}), \text{ and} \\
O(L) &:= \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\mathcal{G}_R}(L) \cup \mathrm{pa}^{\mathcal{G}_R}(L)\}),
\end{aligned}
$$

respectively. All such sets contain at most one node and by slight abuse of notation, we use the same letters to refer to the nodes. We will henceforth suppress the dependence on L if the choice is clear from the context. Figure 4.1 visualizes the above sets.

Now partition $\mathcal{T}_p \setminus \{\mathcal{G}\}$ into the three following disjoint partitions for which there exists a reduced graph sink node $L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}})$ such that $W(L)$, $Y(L)$ and $Z(L)$ is non-empty, respectively. That is, we define

$$
\begin{aligned}
\mathcal{T}_p(\mathcal{G}, W) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } W(L) \neq \emptyset\}, \\
\mathcal{T}_p(\mathcal{G}, Y) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } Y(L) \neq \emptyset\} \setminus \mathcal{T}_p(\mathcal{G}, W), \\
\mathcal{T}_p(\mathcal{G}, Z) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } Z(L) \neq \emptyset\} \\
&\qquad \setminus (\mathcal{T}_p(\mathcal{G}, W) \cup \mathcal{T}_p(\mathcal{G}, Y)).
\end{aligned}
$$

Using that $\mathcal{T}_p(\mathcal{G}, W) \cup \mathcal{T}_p(\mathcal{G}, Y) \cup \mathcal{T}_p(\mathcal{G}, Z) = \mathcal{T}_p(\mathcal{G})$, we can now find a lower bound for the score gap that holds uniformly over all alternative directed tree graphs $\mathcal{T}_p \setminus \{\mathcal{G}\}$:

$$
\begin{aligned}
&\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \\
&= \min \Bigg\{ \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}), \\
&\qquad\qquad \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}), \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \Bigg\}.
\end{aligned}
$$

We now turn to each of these three terms individually and first consider alternative graphs in the partitioning $\mathcal{T}_p(\mathcal{G}, Z)$. The following lower bound consists of possibly highly non-localized conditional dependence properties of observable distribution P_X.

Lemma C.7. *Let $\Pi_Z(\mathcal{G})$ denote all tuples $(z, l, o) \in V^3$ of adjacent nodes $(z \to l) \in \mathcal{E}$ for which there exists a node $o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z, l\}$. It holds that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(z,l,o) \in \Pi_Z(\mathcal{G})} I(X_z; X_o | X_l).$$

The next result proves a lower bound that holds uniformly over all alternative graphs in $\mathcal{T}_p(\mathcal{G}, W)$. The lower bound consists only of local conditional dependence properties. That is, for any subgraph of the causal graph $\mathcal{G}$ of the form $X_o \to X_w \to X_l$ or $X_o \leftarrow X_w \to X_l$ we measure, by means of conditional mutual information, the conditional dependence of the two adjacent nodes X_w and X_l conditional on X_o, $I(X_w; X_l | X_o)$. The lower bound consists of the smallest of all such local conditional dependence measures.

Lemma C.8. *Let $\Pi_W(\mathcal{G})$ denote all tuples $(w, l, o) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ and $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$. It holds that that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l | X_o).$$

A uniform lower bound of the score gap over all alternative graphs in the final partition $\mathcal{T}_p(\mathcal{G}, Y)$ is given by the smallest edge-reversal of any edge in the causal graph $\mathcal{G}$.

Lemma C.9. *It holds that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(j \to i) \in \mathcal{E}} \Delta \ell_{\mathrm{E}}(j \xleftarrow{\cdot\,\cdot} i).$$

An immediate consequence of Lemmas C.7 to C.9 is that the entropy identifiability gap is given by the smallest of the lower bounds derived for each partition, see Theorem 4.5. Thus, it only remains to prove Lemmas C.7 to C.9.

Proof of Lemma C.7: Let $\tilde{\mathcal{G}} \in \Pi_Z(\mathcal{G})$ such that $Z \neq \emptyset$. This implies that $Y = W = \emptyset$ as L can only have one parent in $\mathcal{G}$. Furthermore, $D = \emptyset$ as L can only have one parent in $\tilde{\mathcal{G}}$ and $O \neq \emptyset$ for otherwise L would have been deleted by the deletion procedure in Section 4.5. Assume without loss of generality that $O = \{O_1, \ldots, O_k\}$ for some $k \in \mathbb{N}$. The two subgraphs are illustrated in Figure C.6. For ease of notation, fix any $1 \leq i \leq k$ and denote $O := O_i$. We note that in $\tilde{\mathcal{G}}$ the following d-separation holds

$$Z \perp\!\!\!\perp_{\tilde{\mathcal{G}}} O \,|\, L.$$

Thus, we have for all probability measures $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ over nodes V it holds that $Z \perp\!\!\!\perp O \,|\, L$ as the path between Z and O is blocked by L and all

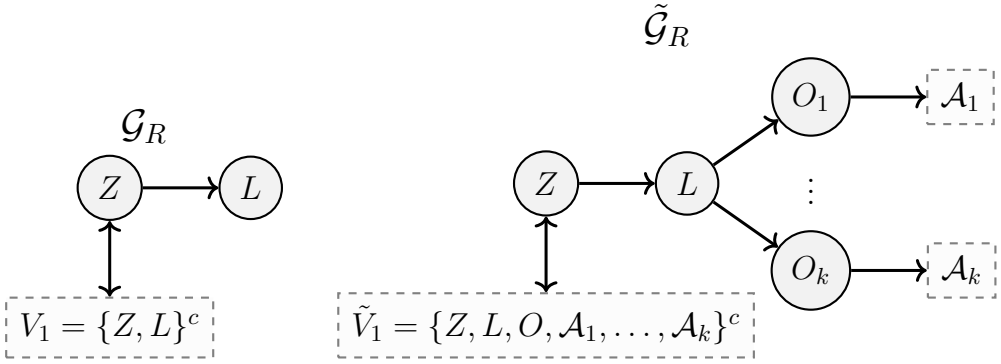


Figure C.6: Illustration of the reduced form graphs $\mathcal{G}_R$ and $\tilde{\mathcal{G}}_R$ for the case $\tilde{\mathcal{G}} \in \Pi_Z(\mathcal{G})$. $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are possibly empty sets of nodes, and dashed rectangle nodes denotes a possibly multi-node subgraph over the variables enclosed. The bi-directed edges means that the edge can be directed in both directions. An edge pointing into the multi-node subgraph, can possibly be multiple edges into distinct nodes of the subgraph.

probability measures generated in accordance with an SCM are automatically Markovian with respect to the generating graph $\tilde{\mathcal{G}}$. Recall from Lemma 4.2 that

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\mathrm{KL}}(P_X \| Q) \\ &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X).\end{aligned}$$

Now fix $Q = q \cdot \lambda^p \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ and note that it factorizes as $Q = Q_{A|Z,O,L} Q_{Z|L} Q_{O|L} Q_L$, i.e., the density q factorizes as

$$\begin{aligned}q(x) &= q_{A|Z,O,L}(a|z,o,l) q_{Z,O,L}(z,o,l) \\ &= q_{A|Z,O,L}(a|z,o,l) q_{Z|L}(z|l) q_{O|L}(o|l) q_L(l),\end{aligned}$$

for λ^p-almost all $x = (a,z,o,l) \in \mathbb{R}^p$ where $A = V \setminus \{Z, O, L\}$. Hence, the cross entropy is splits additively into

$$\begin{aligned}h(P_X, Q) \geq\ & \mathbb{E}[-\log(q_{A|Z,O,L}(A|Z,O,L))] \\ &+ \mathbb{E}[-\log(q_{Z|L}(Z|L))] \\ &+ \mathbb{E}[-\log(q_{O|L}(O|L))] \\ &+ \mathbb{E}[-\log(q_L(L))]. \end{aligned} \tag{C.25}$$

Now note, e.g., that for a conditional distribution (Markov kernel) $Q_{Z|L}$ it holds that

$$\begin{aligned}0 \leq D_{\mathrm{KL}}(P_{Z|L} P_L \| Q_{Z|L} P_L) &= \mathbb{E}\left[-\log\left(\frac{q_{Z|L}(Z|L) p_L(L)}{p_{Z|L}(Z|L) p_L(L)}\right)\right] \\ &= \mathbb{E}[-\log(q_{Z|L}(Z|L))] - \mathbb{E}[-\log(p_{Z|L}(Z|L))],\end{aligned}$$

proving that

$$\mathbb{E}[-\log(q_{Z|L}(Z|L))] \geq \mathbb{E}[-\log(p_{Z|L}(Z|L))].$$

By similar arguments, we get that the three other terms in the lower bound of Equation (C.25) is bounded below by

$$\begin{aligned}\mathbb{E}[-\log(q_{A|Z,O,L}(A|Z,O,L))] &\geq \mathbb{E}[-\log(p_{A|Z,O,L}(A|Z,O,L))],\\ \mathbb{E}[-\log(q_{O|L}(O|L))] &\geq \mathbb{E}[-\log(p_{O|L}(O|L))],\\ \mathbb{E}[-\log(q_L(L))] &\geq \mathbb{E}[-\log(p_L(L))].\end{aligned}$$

This implies that

$$\inf_{Q\in\{\tilde{\mathcal{G}}\}\times\mathcal{F}(\tilde{\mathcal{G}})\times\mathcal{P}^p} h(P_X, Q) \geq h(P_X, Q^*),$$

where $Q^* = P_{A|Z,O,L}P_{Z|L}P_{O|L}P_L$. On the other hand, we know that P_X factorizes as $P_X = P_{A|Z,O,L}P_{Z,O|L}P_L$. Thus we have the following entropy score gap lower bound

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq h(P_X, Q^*) - h(P_X)\\ &= D_{\mathrm{KL}}(P_X \| Q^*)\\ &= D_{\mathrm{KL}}(P_{A|Z,O,L}P_{Z,O|L}P_L \| P_{A|Z,O,L}P_{Z|L}P_{O|L}P_L)\\ &= D_{\mathrm{KL}}(P_{Z,O|L}P_L \| P_{Z|L}P_{O|L}P_L)\\ &= D_{\mathrm{KL}}(P_{Z,O|L} \| P_{Z|L}P_{O|L}|P_L)\\ &= I(Z;O|L).\end{aligned}$$

Let $\Pi_Z(\mathcal{G})$ denote all tuples $(z,l,o) \in V^3$ of adjacent nodes $(z \to l) \in \mathcal{E}$ for which there exists a node $o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z,l\}$. For any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)$ we can, by the above considerations, find a tuple $(z,l,o) \in \Pi_Z(\mathcal{G})$ such that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq I(X_o; X_z \,|\, X_l).$$

We conclude that

$$\min_{\tilde{\mathcal{G}}\in\mathcal{T}_p(\mathcal{G},Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(z,l,o)\in\Pi_Z(\mathcal{G})} I(X_o; X_z|X_l).$$

□

Proof of Lemma C.8: Fix any $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$ and L with $W \neq \emptyset$ such that $Z = Y = \emptyset$. We have illustrated the subgraph $\mathcal{G}_R$ in Figure C.7 and the possible subgraphs $\tilde{\mathcal{G}}_R$ in Figure C.8. .

Note that for any of the three possible local graph structures presented in Figure C.8 there exists an $A \in \{O_1, \ldots, O_k, D\}$ such that $L \perp\!\!\!\perp_{\tilde{\mathcal{G}}_R} W \,|\, A$, i.e., A blocks the path between L and W. Thus, for all probability measures $Q \in \{\tilde{\mathcal{G}}\}\times\mathcal{F}(\tilde{\mathcal{G}})\times\mathcal{P}^p$

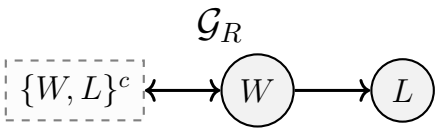


Figure C.7: Illustrations of the $\mathcal{G}_R$ subgraph for for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$.

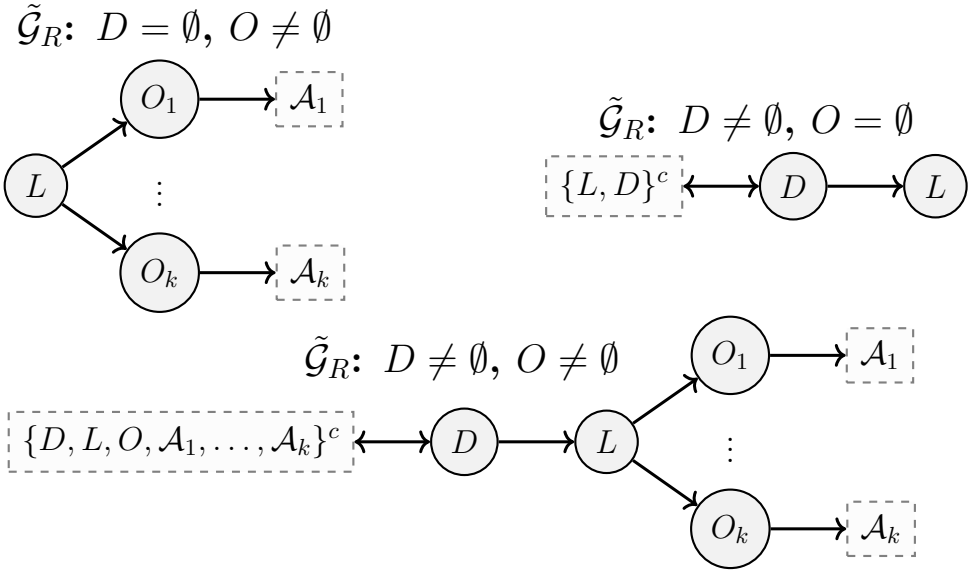


Figure C.8: Illustrations of the possible $\tilde{\mathcal{G}}_R$ subgraphs for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$.

over nodes $V = \{1, .., p\}$ it always holds that $L \perp\!\!\!\perp W \mid A$. By arguments similar to those in the proof of Lemma C.7, we note that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X).$$

and that

$$\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) \geq h(P_X, Q^*),$$

for $P_X = P_{K|W,L,A} P_{W,L|A} P_A$ and $Q^* = P_{K|W,L,A} P_{L|A} P_{W|A} P_A$ where $K = V \setminus \{W, L, A\}$. To that end, we now have that

$$\begin{aligned}
\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq h(P_X, Q^*) - h(P_X) \\
&= D_{\mathrm{KL}}(P_X \| Q^*) \\
&= D_{\mathrm{KL}}(P_{K|W,L,A} P_{W,L|A} P_A \| P_{K|W,L,A} P_{L|A} P_{W|A} P_A) \\
&= D_{\mathrm{KL}}(P_{W,L|A} P_A \| P_{L|A} P_{W|A} P_A) \\
&= D_{\mathrm{KL}}(P_{W,L|A} \| P_{L|A} P_{W|A} | P_A) \\
&= I(W; L | A).
\end{aligned}$$

Let $\hat{\Pi}_W(\mathcal{G})$ denote all tuples $(w, l, a) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ for which there exists a node $a \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{w\}$. Now note that for any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$

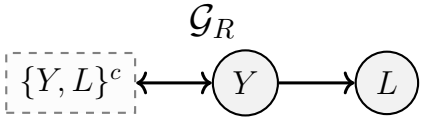


Figure C.9: Illustrations of the $\mathcal{G}_R$ subgraph for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$.

we can, by the above considerations, find a tuple $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ such that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq I(X_w; X_l \,|\, X_a). \tag{C.26}$$

Conversely for any tuple $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ we can construct a graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$ such that Equation (C.26) holds. To see this, fix $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ and construct $\tilde{\mathcal{G}}$ such that the subtree with root node l is identical in both $\mathcal{G}$ and $\tilde{\mathcal{G}}$ and a blocks the path between l and w in $\tilde{\mathcal{G}}$. Therefore, the following lower bound holds and it is not unnecessarily small.

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,a) \in \hat{\Pi}_W(\mathcal{G})} I(X_w; X_l \,|\, X_a).$$

For any $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ it either holds that $a \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$ or that there exists an $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$ blocking the path between a and l such that $X_l \perp\!\!\!\perp X_a | X_o$. Furthermore, we note that as $X_l \perp\!\!\!\perp (X_o, X_a) \,|\, X_w$ we have that

$$\begin{aligned}
I(X_w; X_l | X_a) &= h(X_l | X_a) - h(X_l | X_a, X_w) \\
&= h(X_l | X_a) - h(X_l | X_w) \\
&= h(X_l | X_a) - h(X_l | X_o, X_w) \\
&\geq h(X_l | X_a, X_o) - h(X_l | X_o, X_w) \\
&= h(X_l | X_o) - h(X_l | X_o, X_w) \\
&= I(X_w; X_l \,|\, X_o),
\end{aligned}$$

as further conditioning reduces conditional entropy. Let $\Pi_W(\mathcal{G})$ denote all tuples $(w, l, o) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ and $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$. By the above considerations we conclude that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o).$$

□

Proof of Lemma C.9: Fix $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$ and L such that $Y \neq \emptyset$. It holds that $W = Z = \emptyset$. We have illustrated the $\mathcal{G}_R$ in Figure C.9 and the three possible subgraphs $\tilde{\mathcal{G}}_R$ in Figure C.10.

Note that for any of the three possible local graph structures of $\tilde{\mathcal{G}}_R$ illustrated in Figure C.10 we have that for all probability measures $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$

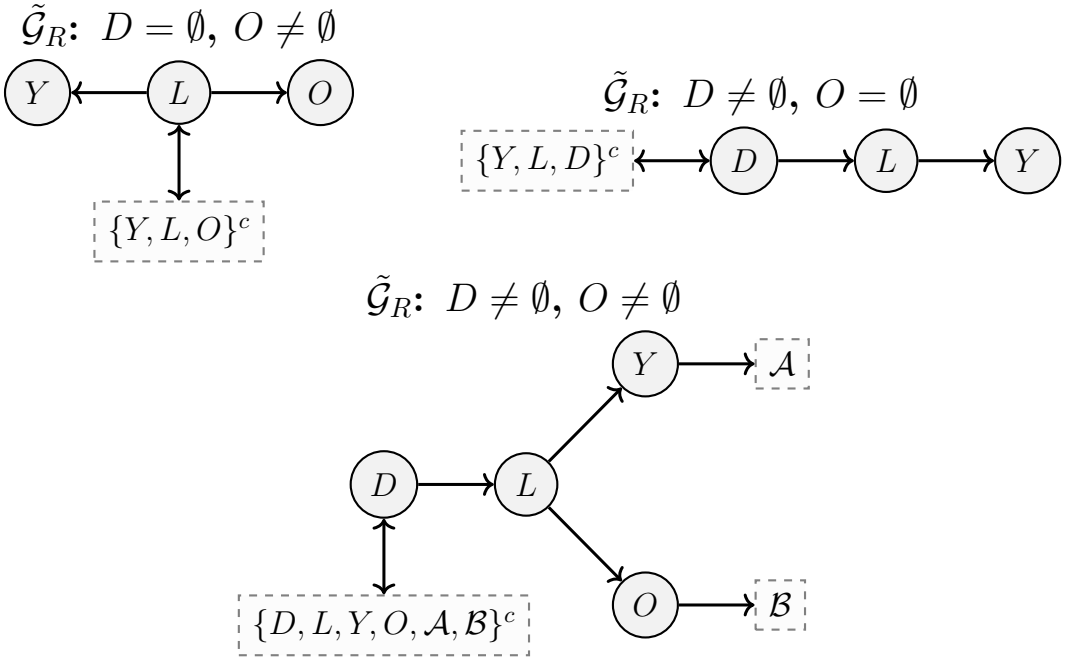


Figure C.10: Illustrations of the possible $\tilde{\mathcal{G}}_R$ subgraphs for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$.

factorizes as $Q_{A|L,Y}Q_{L,Y}$, where $A = V \setminus \{L, Y\}$. It always holds that $Q_{L,Y}$ is the simultaneous distributions of $(\tilde{L}, \tilde{Y})$ generated in accordance with a structural equation model of the form

$$\tilde{Y} := \tilde{f}_Y(\tilde{L}) + \tilde{N}_Y, \tag{C.27}$$

where $\tilde{f}_Y(l) = \mathbb{E}[Y|L = l]$ for all $l \in \mathbb{R}$, and any $\mathcal{L}(\tilde{N}_Y), \mathcal{L}(\tilde{L}) \in \mathcal{P}$ with $\tilde{N}_Y \perp\!\!\!\perp \tilde{L}$. Now recall that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X),$$

and notice that by arguments similar to those in the proof of Lemma C.7 we get

$$\begin{aligned} h(P_X, Q) &= h(P_X, Q_{A|L,Y}Q_{L,Y}) \\ &= \mathbb{E}[-\log(q_{A|L,Y}(A|L, Y))] + h(P_{L,Y}, Q_{L,Y}) \\ &\geq \mathbb{E}[-\log(p_{A|L,Y}(A|L, Y))] + h(P_{L,Y}, Q_{L,Y}), \end{aligned}$$

and that $h(P_X) = \mathbb{E}[-\log(p_{A|L,Y}(A|L, Y)] + h(P_{L,Y})$. Thus, we have that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_{L,Y}, Q_{L,Y}) - h(P_{L,Y}).$$

For any $Q = Q_{A|L,Y}Q_{L,Y} \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ we have that $Q_{L,Y}$ is uniquely determined by a marginal distribution $Q_L \in \mathcal{P}$ and the noise distribution of

$\tilde{N}_Y \sim q_{\tilde{N}_Y} \cdot \lambda \in \mathcal{P}$ from the additive noise structural assignment in Equation (C.27) for $\tilde{Y}$. Thus, the density $q_{L,Y}$ of $Q_{L,Y}$ is given by

$$q_{L,Y}(l,y) = q_{Y|L}(y|l)q_L(l) = q_{\tilde{N}_Y}(y - \tilde{f}_Y(l))q_L(l) = q_{\tilde{N}_Y}(y - \mathbb{E}[Y|L=l])q_L(l).$$

Hence,

$$\begin{aligned} h(P_{L,Y}, Q_{L,Y}) &= \mathbb{E}\left[-\log\left(q_{L,Y}(L,Y)\right)\right] \\ &= \mathbb{E}\left[-\log\left(q_{Y|L}(Y|L)\right)\right] + \mathbb{E}\left[-\log\left(q_L(L)\right)\right] \\ &= \mathbb{E}\left[-\log\left(q_{\tilde{N}_Y}(Y - \mathbb{E}[Y|L])\right)\right] + h(P_L, Q_L) \\ &= h(Y - \mathbb{E}[Y|L], \tilde{N}_Y) + h(P_L, Q_L) \\ &\geq h(Y - \mathbb{E}[Y|L]) + h(L), \end{aligned}$$

where we used that $h(P,Q) = D_{\mathrm{KL}}(P,Q) + h(P) \geq h(P)$. Thus, we have that

$$\begin{aligned} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_{L,Y}, Q_{L,Y}) - h(P_{L,Y}) \\ &\geq h(Y - \mathbb{E}[Y|L]) + h(L) - h(L - \mathbb{E}[L|Y]) - h(Y) \\ &= \Delta\ell_{\mathrm{E}}(Y \xleftarrow{\cdot\,\cdot} L). \end{aligned}$$

We conclude that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G},Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(i \to j) \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{\cdot\,\cdot} i).$$

□

C.4.4.3. Remaining proof of Section 4.5

Proof of Theorem 4.6:

Consider a graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)$ and let $\mathcal{G}_{R,1} = (\mathcal{E}_{R,1}, V_{R,1})$ and $\tilde{\mathcal{G}}_{R,1} = (\tilde{\mathcal{E}}_{R,1}, V_{R,1})$ be the reduced graphs after the initial edge and node deletion procedure of Section 4.5.3. The deletion procedure does not change the score gap, that is,

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}).$$

For any $i \geq 1$ and fixed $\mathcal{G}_{R,i}$ and $\tilde{\mathcal{G}}_{R,i}$ we define

$$\mathbb{L}_{R,i} := \{L \in V_{R,i} : \mathrm{ch}^{\mathcal{G}_{R,i}}(L) = \emptyset \wedge (\mathrm{pa}^{\mathcal{G}_{R,i}}(L) \neq \mathrm{pa}^{\mathcal{G}_{R,i}}(L) \vee \mathrm{ch}^{\mathcal{G}_{R,i}}(L) \neq \emptyset)\}.$$

Now fix $L_1 \in \mathbb{L}_{R,1}$ such that $Z_1 \neq \emptyset$, where Y_1, Z_1, W_1, D_1 and O_1 is defined similarly to the variables in Section 4.5. Let $O_1 = \{O_{1,1}, \ldots, O_{1,k_1}\}$, for some $k_1 \in \mathbb{N}$.

Assume that there exists an $i \in \{1, \ldots, k_1\}$ such that $(Z_1 \to O_{1,i}) \in \mathcal{E}_{R,1}$ in which case we have the following two paths in $\mathcal{G}_{R,1}$ and $\tilde{\mathcal{G}}_{R,1}$

$$\mathcal{G}_{R,1} : O_{1,i} \leftarrow Z_1 \to L_1, \quad \text{and} \quad \tilde{\mathcal{G}}_{R,1} : Z_1 \to L_1 \to O_{1,i}.$$

Since $O_{1,i} \perp\!\!\!\perp_{\tilde{\mathcal{G}}_{R,1}} Z_1 \,|\, L_1$ an entropy score gap lower bound is given by

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) \geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,1}) \geq I(O_{1,i}; Z_1 | L_1),$$

by infimum cross entropy and factorizing arguments similar to those from the proof of Lemma C.8. Now note that $(Z_1, O_{1,i}, L_1) \in \Pi_W(\mathcal{G}_{R,1}) \subseteq \Pi_W(\mathcal{G})$ as $(Z_1 \to O_{1,i}) \in \mathcal{E}_{R,1}$ and $L_1 \in \mathrm{ch}^{\mathcal{G}_{R,1}}(Z_1) \setminus \{O_{1,i}\} \subseteq (\mathrm{ch}^{\mathcal{G}_{R,1}}(Z_1) \setminus \{O_{1,i}\}) \cup \mathrm{pa}^{\mathcal{G}_{R,1}}(Z_1)$. Hence,

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l | X_o). \tag{C.28}$$

Conversely, assume for all $i \in \{1, .., k_1\}$ that $(Z_1 \to O_{1,i}) \notin \mathcal{E}_{R,1}$. Let $\hat{\mathcal{G}}_{R,1} = (\hat{\mathcal{E}}_{R,1}, V_{R,1})$ denote an intermediate graph where $\hat{\mathcal{E}}_{R,1}$ is identical to $\tilde{\mathcal{E}}_{R,1}$ except the edges $\{(L_1 \to O_{1,i}) : 1 \leq i \leq k_1\} \subseteq \tilde{\mathcal{E}}_{R,1}$ is exchanged for the edges $\{(Z_1 \to O_{1,i}) : 1 \leq i \leq k_1\}$. It holds that

$$\begin{aligned}\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) &= \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) + \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) \\ &\geq \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}).\end{aligned}$$

Note that this score gap lower bound is still strictly positive as $\hat{\mathcal{G}}_{R,1} \neq \mathcal{G}_{R,1}$. To realize the last inequality, simply note that as $O_{1,i} \perp\!\!\!\perp L_1 \,|\, Z_1$ we have for all $i \in \{1, \ldots, k_1\}$ that

$$\begin{aligned}2\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}, O_{1,i}) &= \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|L_1])^2] \\ &\geq \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|Z_1, L_1])^2] \\ &= \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|Z_1])^2] \\ &= 2\ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}, O_{1,i}).\end{aligned} \tag{C.29}$$

Now since all edges in $\tilde{\mathcal{G}}_{R,1}$ and $\hat{\mathcal{G}}_{R,1}$ coincide except the incoming edges into $O_{1,1}, \ldots, O_{1,k_1}$ we get that

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) = \sum_{i=1}^{k_1} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, O_{1,i}) - \ell_{\mathrm{G}}(\mathcal{G}, O_{1,i}) \geq 0,$$

where the inequality follows from Equation (C.29). Now both $\hat{\mathcal{G}}_{R,1}$ and $\mathcal{G}_{R,1}$ have a childless node L_1 with the same parent Z_1 so we let $\tilde{\mathcal{G}}_{R,2}$ and $\mathcal{G}_{R,2}$ denote these two graphs where the node L_1 and its incoming edge are deleted. This deletion does not change the graph scores, i.e.,

$$\ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) = \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,2}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,2}).$$

Now fix $L_2 \in \mathbb{L}_{R,2}$ and define Y_2, Z_2, W_2, D_2 and $O_2 = \{O_{2,1}, \ldots, O_{2,k_2}\}$ accordingly.

If either Y_2 or W_2 is non-empty, we use the score gap lower bound previously discussed in Lemma C.8 and Lemma C.9. If Z_2 is non-empty, we can repeat the above procedure and iteratively move edges and delete nodes until the we arrive at the first $i \in \mathbb{N}$ with $\tilde{\mathcal{G}}_{R,i}$ and $\mathcal{G}_{R,i}$ being the iteratively reduced graphs and $L_{R,i} \in \mathbb{L}_{R,i}$ where either

i) Y_i or W_i is non-empty and we get that $\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G})$ is lower bounded by a bound similar to the form of Lemma C.8 or Lemma C.9. That is,

$$\begin{aligned}\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,i}) &\geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,i}) \\ &\geq \min\Bigg\{ \min_{j \to i \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(i \xleftrightarrow{} j) \\ &\qquad , \min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o)\Bigg\}.\end{aligned}$$

ii) Z_i is non-empty and there exists a $j \in \{1, \ldots, k_i\}$ such that $(Z_i \to O_{i,j}) \in \mathcal{G}_{R,i}$. As previously argued, the score gap lower bound of Equation (C.28) applies. That is

$$\begin{aligned}\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,i}) &\geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,i}) \\ &\geq \min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o).\end{aligned}$$

Note that whenever we do not meet scenario i) or ii) we remove a node in both graphs that is a sink node in the reduced true causal graph $\mathcal{G}_{R,i}$ and the intermediate graph $\hat{\mathcal{G}}_{R,i}$. After at most $p-2$ graph reduction iterations of not encountering scenario i) or ii) we are left with two different graphs on two nodes, in which case the score gap is an edge reversal. We conclude that

$$\begin{aligned}\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) &\geq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,i}) \\ &\geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,i}) \\ &\geq \min\left\{ \min_{i \to j \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftrightarrow{} i), \min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o)\right\}.\end{aligned}$$

□

Bibliography

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 1576–1584, 2015.

Acemoglu, D., Johnson, S., and Robinson, J. A. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5): 1369–401, 2001. DOI: 10.1257/aer.91.5.1369.

Albouy, D. Y. The colonial origins of comparative development: an empirical investigation: comment. *American economic review*, 102(6):3059–76, 2012. DOI: 10.1257/aer.102.6.3059.

Aldrich, J. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989. DOI: 10.1093/oxfordjournals.oep.a041889.

Amemiya, T. *Advanced Econometrics*. Harvard University Press, Cambridge, MA, 1985. DOI: 10.2307/2554459.

Amemiya, T. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2:105–110, 1974. DOI: 10.1016/0304-4076(74)90033-5.

Anderson, T. W. Some recent developments on the distributions of single-equation estimators. In Hildenbrand, W., editor, *Advances in Econometrics*, page 109–22. Cambridge University Press, Cambridge, UK, 1983. DOI: 10.1017/cbo9781139052160.004.

Anderson, T. W. and Rubin, H. 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. DOI: 10.1214/aoms/1177730090.

Anderson, T. W. and Rubin, H. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 21:570–82, 1950. DOI: 10.1214/aoms/1177729752.

Andrews, I. and Armstrong, T. B. Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, 8:479–503, 2017. DOI: 10.3982/qe700.

Andrews, I., Stock, J. H., and Sun, L. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–53, 2019. DOI: 10.1146/annurev-economics-080218-025643.

Angrist, J. D. and Krueger, A. B. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106:979–1014, 1991. DOI: 10.2307/2937954.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002. DOI: 10.1023/A:1013689704352.

Bagnell, J. A. Robust supervised learning. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 714–719, 2005.

Bartlett, P. L., Dani, V., Hayes, T., Kakade, S., Rakhlin, A., and Tewari, A. High-probability regret bounds for bandit online linear optimization. In *21st Annual Conference on Learning Theory (COLT)*, 2008.

Basmann, R. L. On the asymptotic distribution of generalized linear estimators. *Econometrica*, 28:97–107, 1960a. DOI: 10.2307/1905296.

Basmann, R. L. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, 55:650–59, 1960b. DOI: 10.1080/01621459.1960.10483365.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000. DOI: 10.1613/jair.731.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 129–136. PMLR, 2010.

Berrett, T. B. and Samworth, R. J. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019. DOI: 10.1093/biomet/asz024.

Berrett, T. B., Grose, D., and Samworth, R. J. CRAN R-package 'IndepTest': Nonparametric independence tests based on entropy estimation, 2018. URL https://cran.r-project.org/web/packages/IndepTest.

Berrett, T. B., Samworth, R. J., and Yuan, M. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288 – 318, 2019. DOI: 10.1214/18-aos1688.

Bühlmann, P., Peters, J., and Ernest, J. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014. DOI: 10.1214/14-aos1260.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009.

Blanchet, J., Kang, Y., Murthy, K., and Zhang, F. Data-driven optimal transport cost selection for distributionally robust optimization. In *2019 Winter Simulation Conference (WSC)*, pages 3740–3751. IEEE, 2019. DOI: 10.1109/wsc40007.2019.9004785.

Bollen, K. A. *Structural Equations with Latent Variables*. John Wiley and Sons, New York, NY, 1989. DOI: 10.1002/9781118619179.

Bongers, S. and Mooij, J. M. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018.

Bongers, S., Peters, J., Schölkopf, B., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics (forthcoming), arXiv preprint arXiv:1611.06221*, 2021.

Bowden, R. J. and Turkington, D. A. *Instrumental Variables*. Econometric Society Monographs. Cambridge University Press, Cambridge, UK, 1985. DOI: 10.1017/ccol0521262410.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004. DOI: 10.1017/cbo9780511804441.

Buckles, K. S. and Hungerman, D. M. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics*, 95:711–24, 2013. DOI: 10.1162/rest_a_00314.

Card, D. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, 1993.

Carey, V., Long, L., and Gentleman, R. Bioconductor R-package 'RBGL', 2021. URL https://www.bioconductor.org/packages/release/bioc/html/RBGL.html.

Caruana, R. Multitask learning. *Machine Learning*, 28:41–75, 1997. DOI: 10.1007/978-1-4615-5529-2_5.

Cayley, A. A theorem on trees. *Quart. J. Math.*, 23:376–378, 1889.

Chamberlain, G. Decision theory applited to an instrumental variables model. *Econometrica*, 75:609–652, 2007. DOI: 10.1111/j.1468-0262.2007.00764.x.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R., and Woutersen, T. An expository note on the existence of moments of fuller and hful estimators. In Baltagi, B. H., Hill, R. C., Newey, W. K., and White, H. L., editors, *Essays in Honor of Jerry Hausman (Advances in Econometrics, Vol. 29)*, pages 87–106. Emerald Group Publishing Limited, Bingley, UK, 2012. DOI: 10.1108/s0731-9053(2012)0000029009.

Chen, H., Wang, Y., Li, R., and Shear, K. A note on a nonparametric regression test through penalized splines. *Statistica Sinica*, 24:1143, 2014. DOI: 10.5705/ss.2012.230.

Chen, X. and Christensen, T. M. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018. DOI: 10.3982/qe722.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–54, 2002.

Chmelarova, V. and Hill, R. C. The hausman pretest estimator. *Economics Letters*, 108:96–9, 2010. DOI: 10.1016/j.econlet.2010.04.027.

Chow, C. K. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968. DOI: 10.1109/tit.1968.1054142.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (forthcoming)*, 2021. DOI: 10.1109/tpami.2021.3094760.

Chu, Y. J. and Liu, T. H. On the shortest arborescence of a directed graphs. *Science Sinica*, 14:1396–1400, 1965.

Claassen, T., Mooij, J. M., and Heskes, T. Learning sparse causal models is not NP-hard. In Nicholson, A. and Smyth, P., editors, *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 172–81. Corvallis, Oregon: AUAI Press, 2013.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 2006. DOI: 10.1002/047174882X.

Cragg, J. G. and Donald, S. G. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9:222–40, 1993. DOI: 10.1017/s0266466600007519.

Csurka, G. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

Danks, D. and Plis, S. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, 2013.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011. DOI: 10.3982/ecta6539.

Daume III, H. and Marcu, D. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006. DOI: 10.1613/jair.1872.

Davidson, R. and MacKinnon, J. G. Confidence sets based on inverting anderson–rubin tests. *The Econometrics Journal*, 17:S39–S58, 2014. DOI: 10.1111/ectj.12015.

Dhrymes, P. *Econometrics: Statistical Foundations and Applications*. Springer, New York, NY, 1974.

Didelez, V., Meng, S., and Sheehan, N. A. Assumptions of iv methods for observational epidemiology. *Statistical Science*, 25:22–40, 2010. DOI: 10.1214/09-sts316.

Dominguez, I. S., Aguirre, A. H., and Diharce, E. V. The Gaussian polytree eda with copula functions and mutations. In *EVOLVE-A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation*, pages 123–153. Springer, Berlin, DE, 2013. DOI: 10.1007/978-3-642-32726-1_3.

Dufour, J.-M. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65:1365–87, 1997. DOI: 10.2307/2171740.

Edmonds, J. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240, 1967. DOI: 10.6028/jres.071b.032.

El Ghaoui, L., Lanckriet, G. R. G., and Natsoulis, G. Robust classification with interval data. Technical report, 2003.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018. DOI: 10.1007/s10107-017-1172-1.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. *Regression: models, methods and applications*. Springer, Berlin, DE, 2013.

Fisher, F. M. *The identification problem in econometrics*. McGraw-Hill, New York, NY, 1966. DOI: 10.2307/2552045.

Fisher, R. A. *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK, 1935. DOI: 10.1136/bmj.1.3923.554-a.

Frisch, R. Statistical versus theoretical relations in economic macrodynamics. *Memorandum for the Business Cycle Conference at Cambridge July 1938 (mimeographed)*, 1938.

Fuller, W. A. Some properties of a modification of the limited information estimator. *Econometrica*, 45:939–53, 1977. DOI: 10.2307/1912683.

Gautier, E., Rose, C., and Tsybakov, A. High-dimensional instrumental variables regression and confidence sets. TSE Working Papers 18-930, 2018.

Goldberger, A. S. Structural equation methods in the social sciences. *Econometrica*, 40:979–1001, 1972. DOI: 10.2307/1913851.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. DOI: 10.2307/1912791.

Greene, W. H. *Econometric analysis*. Pearson Education, Upper Saddle River, NJ, 2003.

Guggenberger, P. The impact of a hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory*, 26:369–382, 2010. DOI: 10.1017/S0266466609100026.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer, Berlin, DE, 2002. DOI: 10.1007/b97848.

Haavelmo, T. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944. DOI: 10.2307/1906935.

Hahn, J. and Hausman, J. A new specification test for the validity of instrumental variables. *Econometrica*, 70:163–89, 2002. DOI: 10.1111/1468-0262.00272.

Hahn, J. and Hausman, J. Estimation with valid and invalid instruments. *Annales d'Économie et de Statistique*, (79/80):25–57, 2005. DOI: 10.2307/20777569.

Hahn, J., Hausman, J., and Kuersteiner, G. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *Econometrics Journal*, 7:272–306, 2004. DOI: 10.1111/j.1368-423x.2004.00131.x.

Hall, A. R. *Generalized method of moments*. Oxford University Press, Oxford, UK, 2005.

Han, Y., Jiao, J., Weissman, T., and Wu, Y. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228 – 3250, 2020. DOI: 10.1214/19-aos1927.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning (ICML)*, pages 1414–1423. PMLR, 2017.

Hastie, T. CRAN R-package 'GAM': Generalized additive models, 2020. URL cran.r-project.org/web/packages/gam/.

Hausman, J. A. Specification tests in econometrics. *Econometrica*, 46:1251–71, 1978. DOI: 10.2307/1913827.

Heinze-Deml, C. and Meinshausen, N. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110:303–348, 2021. DOI: 10.1007/s10994-020-05924-1. (arXiv:1710.11469v5).

Horowitz, J. L. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011. DOI: 10.3982/ECTA8662.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47: 663–85, 1952. DOI: 10.1080/01621459.1952.10483446.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems (NeurIPS)*, 21:689–696, 2008a.

Hoyer, P., Shimizu, S., Kerminen, A., and Palviainen, M. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008b. DOI: 10.1016/j.ijar.2008.02.006.

Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. Technical report, 2013.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–439, 2012.

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., and Danks, D. Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models (PGM)*, 2016.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, 2015. DOI: 10.1017/cbo9781139025751.

Imbens, G. W. and Angrist, J. D. Identification and estimation of local average treatment effects. *Econometrica*, 62:467–75, 1994. ISSN 00129682, 14680262. DOI: 10.3386/t0118.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, 2013.

Jakobsen, M. E. and Peters, J. Distributional Robustness of K-class Estimators and the PULSE. *The Econometrics Journal (forthcoming)*, 2021. DOI: 10.1093/ectj/utab031.

Jakobsen, M. E., Shah, R., Bühlmann, P., and Peters, J. Structure Learning for Directed Trees. *arXiv preprint arXiv:2108.08871*, 2021.

Janzing, D., Peters, J., Mooij, J. M., and Schölkopf, B. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 249–257. AUAI Press, 2009.

Janzing, D., Rubenstein, P. K., and Schölkopf, B. Structural causal models for macro-variables in time-series. *arXiv preprint arXiv:1804.03911*, 2018.

Jorgenson, D. W. and Laffont, J.-J. Efficient estimation of nonlinear simultaneous equations with additive disturbances. In *Annals of Economic and Social Measurement, Volume 3, number 4*, pages 615–640. National Bureau of Economic Research, Cambridge, MA, 1974.

Judge, G. G. and Mittelhammer, R. C. A minimum mean squared error semiparametric combining estimator. In Baltagi, B. H., Hill, R. C., Newey, W. K., and White, H. L., editors, *Essays in Honor of Jerry Hausman (Advances in Econometrics, Vol. 29)*, pages 55–85. Emerald Group Publishing Limited, Bingley, UK, 2012. DOI: 10.1108/s0731-9053(2012)0000029008.

Kadane, J. B. Comparison of k-class estimators when the disturbances are small. *Econometrica*, 39:723–737, 1971. DOI: 10.2307/1909575.

Kalisch, M. and Bühlman, P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Keane, M. P. Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156:3–20, 2010. DOI: 10.1016/j.jeconom.2009.09.003.

Kelejian, H. H. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374, 1971. DOI: 10.1080/01621459.1971.10482270.

Kim, S.-J., Magnani, A., and Boyd, S. Robust fisher discriminant analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 18, pages 659–666, 2006.

Kiviet, J. Testing the impossible: Identifying exclusion restrictions. *Journal of Econometrics*, 218:294–316, 2020. DOI: 10.1016/j.jeconom.2020.04.018.

Kleibergen, F. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70:1781–1803, 2002. DOI: 10.1111/1468-0262.00353.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

Koopmans, T. C., Rubin, H., and Leipnik, R. B. Measuring the equation systems of dynamic economics. In Koopmans, T., editor, *Statistical Inference in Dynamic Economic Models. Cowles Commission monographs*, volume 10, pages 53–237. Hoboken, NJ: John Wiley and Sons, 1950.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. DOI: 10.1016/0196-8858(85)90002-8.

Lauritzen, S. *Graphical Models*. Oxford University Press, New York, NY, 1996.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10846–10856, Red Hook, NY, 2018. Curran Associates Inc.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, pages 1041–1048, Red Hook, NY, 2009. Curran Associates, Inc.

Mariano, R. S. The existence of moments of the ordinary least squares and two-stage least squares estimators. *Econometrica*, 40:643–52, 1972. DOI: 10.2307/1912959.

Mariano, R. S. Some large-concentration-parameter asymptotics for the k-class estimators. *Journal of Econometrics*, 3:171–177, 1975. DOI: 10.1016/0304-4076(75)90045-7.

Mariano, R. S. Simultaneous equation model estimators: Statistical properties and practical implications. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*, chapter 7, pages 122–43. Blackwell Publishing Ltd, Malden, MA, 2001. DOI: 10.1002/9780470996249.ch7.

Marinazzo, D., Pellicoro, M., and Stramaglia, S. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008. DOI: 10.1103/physreve.77.056215.

Marinazzo, D., Liao, W., Chen, H., and Stramaglia, S. Nonlinear connectivity by Granger causality. *NeuroImage*, 58(2):330 – 338, 2011. DOI: 10.1016/j.neuroimage.2010.01.099.

Marsiglietti, A. and Kostina, V. A lower bound on the differential entropy of log-concave random vectors with applications. *Entropy*, 20(3):185, 2018. DOI: 10.3390/e20030185.

McDonald, J. B. The k-class estimators as least variance difference estimators. *Econometrica*, 45:759–63, 1977. DOI: 10.2307/1911689.

Meinshausen, N. Causality from a distributional robustness point of view. In *IEEE Data Science Workshop*, pages 6–10, 2018. DOI: 10.1109/dsw.2018.8439889.

Meinshausen, N., Hauser, A., Mooij, J., Peters, J., Versteeg, P., and Bühlmann, P. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016. DOI: 10.1073/pnas.1510493113.

Meinshausen, N. and Bühlmann, P. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015. DOI: 10.1214/15-aos1325.

Mogensen, S. W. and Hansen, N. R. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020. DOI: 10.1214/19-aos1821.

Mogstad, M. and Wiswall, M. Linearity in Instrumental Variables Estimation: Problems and Solutions. IZA Discussion Paper 5216, 2010.

Mooij, J. M., Janzing, D., and Schölkopf, B. From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2013.

Moreira, M. J. Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, 152:131–40, 2009. DOI: 10.1016/j.jeconom.2009.01.012.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, pages 10–18, 2013.

Nagar, A. L. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27:575–95, 1959. DOI: 10.2307/1909352.

Nandy, P., Hauser, A., and Maathuis, M. H. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A): 3151–3183, 2018. DOI: 10.1214/17-aos1654.

Newey, W. K. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–56, 2013. DOI: 10.1257/aer.103.3.550.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. DOI: 10.1111/1468-0262.00459.

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning linear dags. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 17943–17954, Red hook, NY, 2020. Curran Associates, Inc.

Ouerd, M. *Learning in belief networks and its application to distributed databases.* PhD Thesis, University of Ottawa, Ottawa, Canada, 2000.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345 – 1359, 2010. DOI: 10.1109/tkde.2009.191.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. DOI: 10.1162/089976603321780272.

Pearl, J. The causal foundations of structural equation modeling. In Hoyle, R. H., editor, *Handbook of structural equation modeling*, pages 68–91. The Guilford Press, New York, NY, 2012. DOI: 10.21236/ada557445.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009. DOI: 10.1017/cbo9780511803161.

Peirce, C. S. A theory of probable inference. In Peirce, C. S., editor, *Studies in logic by members of the Johns Hopkins Univ.*, pages 126–81. Little, Brown and Co, Boston, MA, 1883.

Peters, J., Janzing, D., and Schölkopf, B. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2436–2450, 2011. DOI: 10.1109/tpami.2011.71.

Peters, J., Bauer, S., and Pfister, N. Causal models for dynamical systems. *arXiv preprint arXiv:2001.06208*, 2020.

Peters, J., Wainwright, M., et al. Analyzing greedy search strategies in restricted structural causal models (in preparation), 2022.

Peters, J. and Bühlmann, P. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. DOI: 10.1093/biomet/ast043.

Peters, J. and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015. DOI: 10.1162/neco_a_00708.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1): 2009–53, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, oct 2016. DOI: 10.1111/rssb.12167.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

Pfister, N., Bauer, S., and Peters, J. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51): 25405–11, 2019. DOI: 10.1073/pnas.1905688116.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. Stabilizing variable selection and regression. *Annals of Applied Statistics (forthcoming), arXiv preprint arXiv:1911.01850*, 2021.

Polyanskiy, Y. and Wu, Y. *Lecture notes on information theory*. 2019. URL http://people.lids.mit.edu/yp/homepage/.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, Cambridge, MA, 2009. DOI: 10.7551/mitpress/9780262170055.001.0001.

Racine, J. S. and Hayfield, T. *np: Nonparametric Kernel Smoothing Methods for Mixed Data Types*, 2018. URL https://CRAN.R-project.org/package=np. R package version 0.60–10.

Rebane, G. and Pearl, J. The recovery of causal poly-trees from statistical data. In *Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 222–228, Seattle, WA, 1987.

Reichenbach, H. *The direction of time*. University of California Press, Berkeley, CA, 1956. DOI: 10.2307/2183684.

Reisach, A. G., Seiler, C., and Weichwald, S. Beware of the simulated DAG! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

Robins, J. M. A new approach to causal inference in mortality studies with sustained exposure periods — applications to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–512, 1986. DOI: 10.1016/0270-0255(86)90088-6.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018a.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19:1309–42, 2018b.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983. DOI: 10.1093/biomet/70.1.41.

Rothenberg, T. J. Approximating the distributions of econometric estimators and test statistics. In Griliches, Z. and Intriligator, M., editors, *Handbook of econometrics*, volume 2, pages 881–935. North-Holland Publishing Company, Amsterdam, NL, 1984.

Rothenhäusler, D., Ernest, J., and Bühlmann, P. Causal inference in partially linear structural equation models. *Annals of Statistics*, 46(6A):2904–2938, 2018. DOI: 10.1214/17-aos1643.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021. DOI: 10.1111/rssb.12398.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2017.

Rubenstein, P. K., Bongers, S., Mooij, J. M., and Schölkopf, B. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.

Rubin, D. B. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974. DOI: 10.1037/h0037350.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. DOI: 10.1198/016214504000001880.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 2019a. DOI: 10.1126/sciadv.aau4996.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Munoz-Mari, J., Nes, E. H., and Peters, J. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):1–13, December 2019b. DOI: 10.1038/s41467-019-10105-3.

Runge, J., Tibau, X.-A., Bruhns, M., Munoz-Mari, J., and Camps-Valls, G. The causality for climate competition. In Escalante, H. J. and Hadsell, R., editors, *PMLR NeurIPS Competition & Demonstration Track Postproceedings*, volume 123 of *Proceedings of Machine Learning Research*, pages 110–120. PMLR, December 2020. URL `https://causeme.uv.es/`.

Sani, N., Lee, J., and Shpitser, I. Identification and estimation of causal effects defined by shift interventions. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2020.

Saunders, C., Gammerman, A., and Vovk, V. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*. Omnipress, 1998.

Schilling, R. L. *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge, UK, 2017. `DOI: 10.1017/CBO9780511810886`.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–62, New York, NY, 2012. Omnipress.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020. `DOI: 10.1214/19-aos1857`.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10):2003–30, 2006.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2): 227 – 244, 2000. `DOI: 10.1016/s0378-3758(00)00115-4`.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. `DOI: 10.1186/s40537-019-0197-0`.

Silva, E. I. *A unified framework for the analysis and design of networked control systems*. PhD Thesis, University of Newcastle, Callaghan, Australia, 2009.

Simon, H. A. Causal ordering and identifiability. In Hood, W. C. and Koopmans, T., editors, *Studies in Econometric Method. Cowles Commission monographs*, volume 14, pages 49–74. Hoboken, NJ: John Wiley and Sons, 1953. `DOI: 10.1007/978-94-010-9521-1_5`.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

Sloane, N. J. A. The on-line encyclopedia of integer sequences, 2021. URL `https://oeis.org/A003024`. The OEIS Foundation Inc. (2021).

Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In Besnard, P. and Hanks, S., editors, *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 499–506, Montréal, CA, 1995. San Mateo, CA: Morgan Kaufmann.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000. `DOI: 10.7551/mitpress/1754.001.0001`.

Staiger, D. and Stock, J. H. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–86, 1997. `DOI: 10.2307/2171753`.

Stock, J. H. and Yogo, M. Testing for weak instruments in linear iv regression. Technical working paper 284, 2002.

Stock, J. H., Wright, J. H., and Yogo, M. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20:518–29, 2002. `DOI: 10.1198/073500102288618658`.

Stramaglia, S., Wu, G.-R., Pellicoro, M., and Marinazzo, D. Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6):066211, 2012. `DOI: 10.1103/physreve.86.066211`.

Stramaglia, S., Cortes, J. M., and Marinazzo, D. Synergy and redundancy in the Granger causal analysis of dynamical networks. *New Journal of Physics*, 16(10): 105003, 2014. `DOI: 10.1088/1367-2630/16/10/105003`.

Sugiyama, M. and Müller, K. Generalization error estimation under covariate shift. In *Workshop on Information-Based Induction Sciences (IBIS)*, 2005.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, pages 1433 – 1440, Vancouver, CA, 2008.

Tarjan, R. E. Finding optimum branchings. *Networks*, 7(1):25–35, 1977. `DOI: 10.1002/net.3230070103`.

Theil, H. Repeated least squares applied to complete equation systems. *The Hague: central planning bureau (mimeographed)*, 1953.

Theil, H. *Economic forecasts and policy*. North-Holland, Amsterdam, NL, 1958.

Tofigh, A. and Sjölund, E. C++ implementation of Edmonds algorithm, 2007. URL `https://github.com/atofigh/edmonds-alg`.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006. `DOI: 10.1007/s10994-006-6889-7`.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, Berlin, DE, 2009. `DOI: 10.1007/b13794`.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41:436–463, 2013. `DOI: 10.1214/12-aos1080`.

Van der Vaart, A. W. *Asymptotic statistics*. Cambridge university press, Cambridge, UK, 3 edition, 2000. `DOI: 10.1017/CBO9780511802256`.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, UAI '90, page 255–270, Amsterdam, NL, 1990a. Elsevier. ISBN 0444892648.

Verma, T. and Pearl, J. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, Amsterdam, NL, 1990b.

Volpi, R., Morerio, P., Savarese, S., and Murino, V. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5495–5504, 2018. `DOI: 10.1109/cvpr.2018.00576`.

Wang, L. and Tchetgen, E. T. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–50, dec 2018. `DOI: 10.1111/rssb.12262`.

Weichwald, S., Jakobsen, M. E., Mogensen, P. B., Petersen, L., Thams, N., and Varando, G. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In Escalante, H. J. and Hadsell, R., editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36. PMLR, 08–14 Dec 2020.

Wiener, N. The theory of prediction. *Modern Mathematics for Engineers*, 1956.

Wold, H. Causality and econometrics. *Econometrica*, 22:162–77, 1954. `DOI: 10.2307/1907540`.

Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, MA, 2010.

Wright, P. G. *Tariff on animal and vegetable oils*. Macmillan Company, New York, NY, 1928.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

Zhang, J. and Spirtes, P. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, page 632–639, San Francisco, CA, 2002. Morgan Kaufmann.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, page 647–655, Arlington, VA, 2009. AUAI Press.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. DAGs with no tears: Continuous optimization for structure learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31, page 9492–9503, Red hook, NY, 2018. Curran Associates, Inc.