Shimeng Huang

# Causal Inference for Complex Data Structures

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

December 2024

Shimeng Huang
shimeng@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

| | |
|---|---|
| **Thesis title:** | Causal Inference for Complex Data Structures |
| **Supervisor:** | Associate Professor Niklas Pfister<br>University of Copenhagen |
| | Professor Jonas Peters<br>ETH Zurich |
| | Professor Susanne Ditlevsen<br>University of Copenhagen |
| **Assessment Committee:** | Professor Niels Richard Hansen (chair)<br>University of Copenhagen |
| | Professor Stefan Bauer<br>Technical University of Munich |
| | Assistant Professor Sara Magliacane<br>University of Amsterdam |
| **Date of Submission:** | December 8,<br>2024 |
| **Date of Defense:** | Feburary 7,<br>2025 |
| **ISBN:** | 978-87-7125-236-1 |



*This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on 8 December 2024. It was supported by research grant 0069071 from Novo Nordisk Fonden.*

献给我的家人。
*To my family.*

# Preface

This thesis is an accumulation of projects during my time as a PhD student under the supervision of Niklas Pfister at the University of Copenhagen and Jonas Peters, who is now at ETH Zurich. The thesis presents three individual manuscripts, each related to a specific complex data structure revolving around two aspects of causal inference: intervention and invariance. Some differences may exist between each chapter and its corresponding publicly available paper which are indicated in "Contributions and Structure". All typographical or mathematical errors are solely my responsibility.


## Acknowledgments

Just over three years ago, I had an online interview in my home office in Toronto, Canada, where I first met my future supervisors, Niklas Pfister and Jonas Peters. Looking back, I am deeply grateful for the opportunity Niklas and Jonas offered me, who was on the other side of the world with little background in causality. Working with Niklas and Jonas has truly been an inspiring and motivating journey. Thank you for all the insightful discussions, detailed comments, and unwavering support over the past three years.

I would also like to thank my other co-authors, Elisabeth Ailer and Niki Kilbertus for the enjoyable collaboration, and Jack Bowden, whom I had the pleasure of collaborating with during my time as a visiting scientist at Novo Nordisk.

My time in Copenhagen has been a real pleasure because of everyone I met here. I feel fortunate to have had the most friendly and caring officemates, who initiated gatherings and organized birthday parties, who cheered me up when I was down, and with whom I became very close friends. Especially the Boyz-n-the-Hood: Pedja, Matt, and Alex—I truly enjoyed all the time we spent together. I particularly want to thank Pedja for all the meaningful and lighthearted conversations, as well as for introducing me to so many wonderful new friends. Also, thanks to Frederik and Yiqing for our Doppelkopf nights!

I would like to thank my family, who have always given me unconditional love and understanding, never held me back when I wanted to explore, and always welcomed me warmly whenever I came home. In particular, I want to thank my dad, who supported me throughout my studies in Canada, prepared gifts and everything I needed long before I could visit back home, pins me at the top of his chats, and is always there for me.

Lastly, I want to thank Lucas, my Hasi, whom I also met in this beautiful city of Copenhagen. Thank you for always being my biggest cheerleader, for growing with me, for all the wonderful trips we have taken and will take, and for all the care and love you have given me. I feel lucky to have you in my life.

Shimeng Huang
December, 2024

# Abstract

This thesis explores and develops causality-related methodology through three individual projects, each focusing on a specific complex data structure. While the problems investigated are diverse, they all center around the concepts of intervention and (causal) invariance. Chapter 1 introduces these foundational ideas, which pervade the remainder of the thesis. Brief introductions of the data structures addressed in the subsequent chapters are also provided in this chapter.

Chapter 2 dives into the first complex data structure, compositional data, where observations lie on a simplex (i.e., each observation is constrained to sum to one). This project is motivated by microbiome research, in which compositions of microbial strains are typically observed. We examine how to define interpretable statistical targets to quantify the effects of the components on a response when the predictor is compositional in regression or classification problems. We develop non-parametric estimators of these effects based on kernels that are specifically suited for compositional data. Our estimators are evaluated on 33 publicly available microbiome datasets and are shown to achieve comparable or superior performance compared to state-of-the-art machine learning methods.

In Chapter 3, we consider sequential data and introduce a new type of change point, termed causal change points, which indicate changes in the causal mechanism relative to a response variable under appropriate assumptions. We propose methods to detect and localize these change points without requiring prior knowledge of the causal structure in the data. These methods leverage the reverse concept of causal invariance—the property that the conditional distribution of the response, given its parents, remains fixed under interventions that do not directly target the response. We demonstrate our methods using two real-world datasets, one on air quality and the other on macroeconomic policy.

The final chapter, Chapter 4, considers sparse causal effects estimation using two-sample summary statistics, a type of summary-level data commonly used in genetics research. In a two-sample summary statistics setting, one does not have access to individual-level data but only to the marginal associations obtained from two samples: one containing paired observations of instruments and covariates, and the other containing paired observations of instruments and the response. We propose a generalized, two-sample summary statistics version of the test statistic considered in spaceIV [Pfister and Peters, 2022], and prove that our proposed test is uniformly asymptotically level. We apply our method, spaceTSIV, to real proteomic and gene-expression data for discovering possible drug targets for coronary artery disease.

# Sammenfatning

Denne afhandling undersøger og udvikler kausalitetsrelateret metodologi gennem tre selvstændige projekter som hver især fokuserer på en specifik kompleks datastruktur. Selvom de undersøgte problemstillinger er forskellige er de alle bygget op omkring begreberne intervention og (kausal) invarians. Kapitel 1 introducerer disse grundlæggende idéer som er centrale for resten af afhandlingen. I dette kapitel gives der også korte introduktioner til de datastrukturer der behandles i de efterfølgende kapitler.

Kapitel 2 omhandler den første komplekse datastruktur, kompositionelle data, hvor observationerne ligger på en simplex (dvs. hver observation summerer til én). Dette projekt er motiveret af mikrobiomforskning hvor man typisk observerer sammensætninger af microbial strains. Vi undersøger hvordan man definerer fortolkelige statistiske mål for at kvantificere effekten af komponenterne på et respons når prædiktoren er kompositionel i regressions- eller klassifikationsproblemer. Vi udvikler ikke-parametriske estimatorer af disse effekter, baseret på kernels der er særligt egnede til kompositionelle data. Vores estimatorer evalueres på 33 offentligt tilgængelige mikrobiomdatasæt og de viser sig at opnå sammenlignelig eller overlegen præstation i forhold til de bedste eksisterende metoder inden for machine learning.

I Kapitel 3 undersøger vi sekventielle data og introducerer en ny type change points, kaldet causal change points, som indikerer ændringer i den kausale mekanisme for responsvariablen under passende antagelser. Vi foreslår metoder til at detektere og lokalisere disse change points uden at kræve forudgående viden om kausale strukturer i data. Disse metoder udnytter pendanten til kausal invarians – at den betingede fordeling af responsen givet dens forældre forbliver uændret under interventioner der ikke direkte retter sig mod responsen. Vi demonstrerer vores metoder ved hjælp af to virkelige datasæt, det ene om luftkvalitet og det andet om makroøkonomisk politik.

Det sidste kapitel Kapitel 4 omhandler estimation af sparse causal effects ved hjælp af two-sample summary statistics, en type opsummerede statistik der ofte anvendes i genetisk forskning. I et two-sample summary statistics setup har man ikke adgang til individdata men kun til marginale associationer opnået fra to stikprøver: en der indeholder sammenkoblede observationer af instrumenter og kovariater, og en anden der indeholder sammenkoblede observationer af instrumenter og respons. Vi foreslår en generaliseret two-sample summary statistics version af den teststatistik der overvejes i spaceIV [Pfister and Peters, 2022] og beviser at vores foreslåede test har uniformt asymptotisk niveau. Vi anvender vores metode spaceTSIV på virkelige data for proteomics og genekspression med henblik på at opdage mulige angrebspunkter for lægemidler for iskæmisk hjertesygdom.

# Contributions and Structure

This thesis contains one introductory chapter and three main chapters, each of which corresponds to a paper. **Chapter 1** introduces the complex data structures considered in the main chapters, the causal questions of interest, their challenges, and summarizes my and my co-authors' contributions. An overview of the contributions of Chapters 2 to 4 are listed below, along with acronyms of the papers which are used in this thesis.

**Chapter 2** proposes a kernel-based non-parametric regression and classification framework for compositional data, corresponding to the paper:

> [**KernelBiome**][Huang et al., 2023]. S. Huang, E. Ailer, N. Kilbertus, and N. Pfister. Supervised Learning and Model Analysis with Compositional Data. *PLOS Computational Biology*, 19(6):e1011240, 2023.

**Chapter 3** introduces the concept of causal change points (CCPs) and studies the problems of detecting and estimating CCPs. This chapter contains a partial revision of the following paper, including a new real data application and with minor typos and errors corrected. The status of this paper is currently 'reject with resubmission' at Biometrika and we are currently preparing to resubmit:

> [**CausalCP**][Huang et al., 2024a]. S. Huang, J. Peters, and N. Pfister. Causal Change Point Detection and Localization. *arXiv Preprint arXiv:2403.12677*, 2024a.

**Chapter 4** formulates sparse causal effect estimation with instrumental variables under the two-sample summary statistics setting, and corresponds to the following paper which was submitted to the 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025):

> [**SpaceTSIV**][Huang et al., 2024b]. S. Huang, N. Pfister, and J. Bowden. Sparse Causal Effect Estimation Using Two-Sample Summary Statistics in the Presence of Unmeasured Confounding. *arXiv Preprint arXiv:2410.12300*, 2024b.

# Contents

**Preface** iv

**Abstract** vi

**Contributions and Structure** ix

**1 Introduction** 1
- 1.1 Intervention, causal models, and invariance . . . 1
- 1.2 Interventions on a simplex . . . 2
- 1.3 Sequential data with unstable causal mechanisms . . . 4
- 1.4 Causal inference with summary-level data . . . 7

**2 Supervised learning and model analysis with compositional data** 11
- 2.1 Introduction . . . 12
- 2.2 Methods . . . 15
- 2.3 Results . . . 23
- 2.4 Discussion and conclusions . . . 28
- 2.A Details on CFI and CPD . . . 29
- 2.B Details on kernels included in `KernelBiome` . . . 30
- 2.C Details and additional results for experiments . . . 35
- 2.D Additional experiments with simulated data . . . 43
- 2.E Background on kernels . . . 49
- 2.F Proofs . . . 54
- 2.G List of kernels implemented in `KernelBiome` . . . 59

**3 Causal change point detection and localization** 71
- 3.1 Introduction . . . 71
- 3.2 Regression change points and causal change points . . . 73
- 3.3 Causal change point detection . . . 78
- 3.4 Causal change point localization . . . 79
- 3.5 Numerical Experiments . . . 83
- 3.6 Discussion . . . 92
- 3.A Additional examples and details on examples . . . 94
- 3.B Algorithms . . . 95
- 3.C Additional numerical experiments and experiment details . . . 95
- 3.D Proofs . . . 102
- 3.E Auxiliary results . . . 104
- 3.F Chow test . . . 105

**4 Sparse causal effect estimation using two-sample summary statistics in the presence of unmeasured confounding 107**
4.1 Introduction . . . 107
4.2 Reduced form IV model and summary statistics . . . 109
4.3 Estimating sparse causal effects with spaceTSIV . . . 113
4.4 Experiments . . . 116
4.5 Discussion . . . 119
4.A Details of test statistics and test-based estimators . . . 121
4.B Regularity conditions . . . 123
4.C Proofs . . . 123
4.D Additional results . . . 128
4.E Experiment details and additional simulation results . . . 131

**Bibliography 135**

# 1 Introduction

In this thesis, we consider three specific complex data structures, and for each data structure, a causality-related question is investigated. As we will see in the following sections, these research topics are inspired by biological and economic problems. Although the three topics are diverse, they are all centered around the ideas of intervention and causal invariance, which are described in more detail in Section 1.1. The rest of this chapter describes the three specific data structures along with the causal questions, and highlights the solutions I have worked on. Section 1.2 introduces compositional data, where one is interested in the causal effects of intervening on the components. Section 1.3 considers sequential data where causal relationships may change at certain time points. Section 1.4 describes two-sample summary statistics data and how instrumental variable methods can be applied to estimate causal effects under unobserved confounding. At the end of Sections 1.2 to 1.4, we provide concise summaries of the data structures and causal questions in colored boxes.

## 1.1 Intervention, causal models, and invariance

This thesis approaches causality from an interventional perspective. Interventions provide an intuitive way to define causality in the physical world. A well-known quote attributed to Paul Holland and Don Rubin states, "no causation without manipulation" [Holland, 1986].

One prominent example of using intervention to infer causality is randomized controlled trials (RCTs), which remain the gold standard for establishing causality in clinical and biomedical research [e.g., Rubin, 1974]. Due to practical or ethical reasons, RCTs are not always possible, and observational data are often the only data source for answering causal questions. If all relevant confounders are assumed to be observed between a set of covariates and a response, one may pursue an adjustment method such as the frontdoor or backdoor adjustment [Pearl, 2009], generalized adjustment [Perković et al., 2015, 2018, Shpitser et al., 2010], or efficient adjustment [Witte et al., 2020]. When there are potential unobserved confounders, instrumental variable (IV) methods [Angrist and Imbens, 1994] may be employed to infer causal effects, which leverages the randomness in the IVs as pseudo-interventions on the covariates.

From a modeling point of view, interventions also distinguish a causal model from a statistical one: while a statistical model specifies a set of distributions over the observed variables in a system, a causal model specifies mappings of well-defined (observed or hypothetical) interventions on some or all of the variables in the system to distributions over the observed variables. That is, for a sample space $\mathcal{X}$, a statistical model specifies

a set of distributions

$$\mathcal{P} \subseteq \left\{ P \,\middle|\, P \text{ is a probability distribution on } \mathcal{X} \right\}.$$

A causal model extends the statistical model and can be described in the following way [Pfister, 2024]. For a fixed index set $\mathcal{I}$, the set of interventions, a causal model specifies a set of functions with domain $\mathcal{I}$

$$\mathcal{P}_{\mathcal{I}} \subseteq \left\{ f \,\middle|\, f : \mathcal{I} \to \mathcal{P} \right\},$$

such that $f(i)$ is a probability distribution on $\mathcal{X}$ corresponding to the intervention $i \in \mathcal{I}$. The observational distribution can be thought of as corresponding to the "do nothing" intervention.

Over the past decades, many specific causal models have been proposed, such as potential outcome models [Rubin, 1974, 2005, Imbens and Rubin, 2015], graphical causal models [Pearl, 2009], and structural causal models (SCM) [Bollen, 1989, Spirtes et al., 2000, Pearl, 2009, Peters et al., 2017, Bongers et al., 2021], each having pros and cons. A more detailed summary of those approaches can be found in Pfister [2024]. In [**CausalCP**] (Chapter 3) and [**SpaceTSIV**] (Chapter 4) where causal models are employed, we focus on linear SCMs.

The abstract notion of a causal model defined above, $\mathcal{P}_{\mathcal{I}}$, imposes no constraints on how interventions are mapped to joint distributions of the observed variables, or on the relationship between the observational and interventional distributions. In many applications, however, it is natural to introduce certain constraints on these mappings. One particular example is the assumption of modularity, or invariance [e.g., Haavelmo, 1943, Aldrich, 1989], which posits that interventions are local, meaning they affect only a subset of variables while leaving the rest of the system unchanged. This means that parts of an interventional distribution can resemble parts of the observational distribution. SCMs formalize this intuition by postulating that the interventional distribution under $\mathrm{do}(X = x)$ is the distribution induced by the SCM in which the structural assignment for $X$ is replaced by $x$ and the rest of the model remains unaltered [Peters et al., 2017].

Causal invariance allows us to discover (part of) the causal mechanism in a system and enables us to achieve robustness and generalization in a statistical model. This includes invariant causal prediction [Peters et al., 2016] and its sequential counterpart [Pfister et al., 2019], which use heterogeneous environments (or time intervals) to discover the causal parents of a response, assuming causal sufficiency. Anchor regression [Rothenhäusler et al., 2021] on the other hand, proposes a loss function that balances between prediction accuracy and invariance, allowing hidden confounding between the covariates and the response. A nice summary of the above and related works is given by Bühlmann [2020].

## 1.2 Interventions on a simplex

The description of a causal model in Section 1.1 advocates an explicit treatment of interventions. This is an aspect of causal models that is often treated implicitly in

causal methodology research and applications, especially under Pearlian graphical models [also known as graphical causal models, Pearl, 2009], as discussed by Dawid [2002]. [**KernelBiome**] (Chapter 2) provides an example where defining interventions is not as straightforward as one might think. In this case, the difficulty arises from the nature of the predictor, which is compositional, such that one cannot intervene on one component without simultaneously intervening on the other components.

Compositional data is a data structure that commonly occurs in geology, ecology, and microbiome studies, where the proportions of a collection of components are measured. A $p$-dimensional observation in compositional data can be represented by a point on a simplex

$$\mathbb{S}^{p-1} := \{x \in [0,1]^p \mid \sum_{j=1}^{p} x^j = 1\}.$$

In many applications, an associated response, such as a disease indicator, is also measured. A real example of such data is given in Figure 1. As in this dataset, compositional data in microbiome studies is often very sparse and high-dimensional. Two common objectives of these applications are: 1) use the compositional predictor to predict the response, and 2) infer how changes in the composition affect the response.

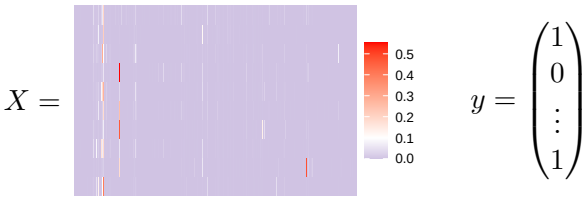



Figure 1: Heatmap of microbiome composition, where each row represents a subject and each column corresponds to a bacterial species, along with an associated binary response variable indicating whether the person is cirrhotic.

The sum-to-one constraint of compositional data poses challenges to statistical methods that consider data in the Euclidean space, as this constraint induces non-trivial dependencies between the components. As we shall see in [**KernelBiome**], ignoring this simplex constraint can lead to wrong conclusions. A popular method in compositional data analysis is the log-contrast model by Aitchison and Bacon-Shone [1984], which is easy to fit and interpret. However, it usually requires adding a small (arbitrary) positive number to the large number of zero components so that the compositional data can be log-transformed. This method is also not capable of handling complex signals or including prior knowledge of the relations between the components.

In [**KernelBiome**], we consider estimating the conditional mean of the response based on kernels, due to their flexibility to capture complex signals, the availability of kernels that are targeted to the simplex, and the possibility of integrating prior knowledge. The log-contrast model can also be shown to be a special case of our method. Moreover, the estimated embeddings can be used for post-analyses that take the compositionality into account, such as visualizing the observations on a lower dimensional space, and

measuring diversity in the compositional samples due to the connection between kernels and distances.

For interpreting the contribution of a component, we start from interpretable targets of inference by defining two specific interventions on the simplex—the space in which compositional data lies—and, based on these interventions, we can properly describe each component's "contribution" to a response. From this angle, we can see that interventions are also connected to variable importance, in the sense that the importance of a variable (or component) can be assessed by how much it affects the response upon an intervention.

It is worth noting that the targets of inference considered in [**KernelBiome**] are by themselves purely observational, and any causal interpretations may require additional assumptions and adaptations. The specific interventions considered in this work are further generalized by Lundborg and Pfister [2023] to arbitrary, well-defined interventions on the simplex, called "perturbations", to distinguish them from the usual concept of intervention on the data-generating mechanism. Their use in causal inference is also discussed therein.

We developed a python library `KernelBiome` in this work which aims to provide a user-friendly framework for regression and classification tasks using compositional features. As a future work, we can extend the current framework to combine both compositional and other metadata features using multi-kernel learning [Gönen and Alpaydın, 2011][1].

**Data structure:** Identically and independently distributed (i.i.d) $n$ observations $\{X_i, Y_i\}_{i=1}^n$ of a random variable $(X, Y)$ where $X \in \mathbb{S}^{p-1}$ is a compositional predictor and $Y \in \mathbb{R}$ is a univariate response.

**Causal question:** How to define an intervention on a simplex? How to estimate the effect of an intervention on a simplex to a response variable?

## 1.3 Sequential data with unstable causal mechanisms

We discussed at the end of Section 1.1 that causal invariance allows us to obtain certain desirable properties. If the response variable was directly intervened on, the assumption of causal invariance no longer holds. This motivates the problem studied in [**CausalCP**] (Chapter 3), if we have data collected sequentially, can we find out whether and when causal invariance does not hold?

The problem we are looking at is the reverse of invariance, where, under certain causal assumptions, we hope to detect and localize changes in the causal mechanism of the response variable without knowing the causal structure, given data observed over time. Figure 2 illustrates some examples of this type of change point, based on causal diagrams.

[1] Thanks to a question in the GitHub repository https://github.com/shimenghuang/KernelBiome/issues/2.

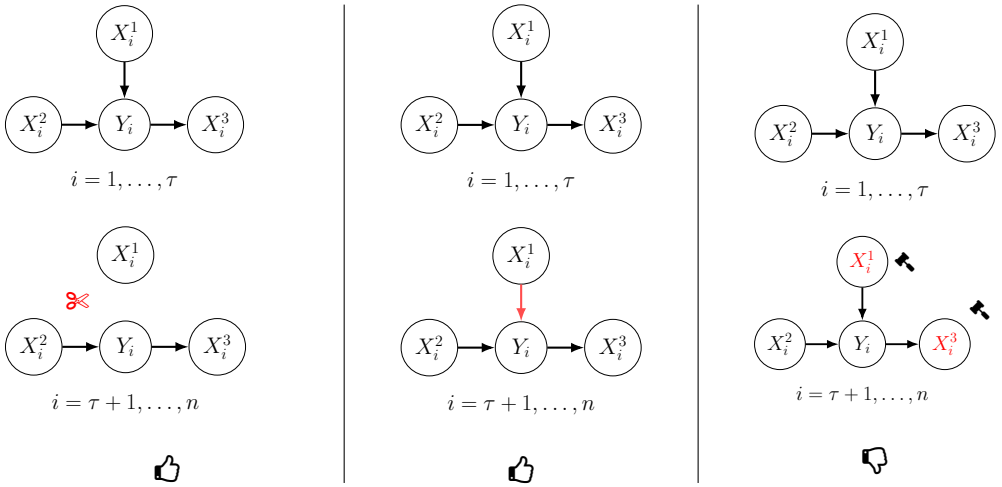



Figure 2: Two examples where the causal mechanism of $Y$ changes (left and middle): one causal parent of $Y$ is removed, the relationship between $Y$ and one of its causal parent changes changes (indicated by the color change of the arrow). One example where the causal mechanism of $Y$ does not change (right): the distributions of $X^1$ and $X^3$ are modified.

Change point detection and localization have been studied extensively, especially in the econometrics literature, spanning from univariate changes, multivariate changes, structural changes, and also in various aspects such as changes in mean, variance (covariance), or the regression coefficients, as well as different types, online or offline. Reviews can be found in Niu et al. [2016], Truong et al. [2020], and Bardet et al. [2020].

One of the main contributions of [**CausalCP**] is to introduce a new type of change point, termed causal change point (CCP), which has not been discussed previously. The closest concept in the literature is "structural change points", referring to changes in the conditional distribution of $Y$ given all covariates. In [**CausalCP**], we refer to structural change points based on linear models as "regression change points" (RCPs), and CCPs form a subset of RCPs.

Although we named this new type of change point "causal change points", its definition is also purely observational and does not refer to an underlying causal model. The causal name stems from the motivating idea of causal invariance, as well as its causal meaning when additional assumptions are satisfied. In [**CausalCP**], we discuss the causal meaning of CCP in the context of SCMs. A general form of sequential SCM is given in Definition 1.3.1, where the key difference compared to a usual SCM is that the function generating a variable also depends on the time index.

**Definition 1.3.1** (Sequential SCM)**.** Let $n \in \mathbb{N}_+$ and assume that for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$, there exists a function $f_i^j : \mathbb{R}^{|\mathrm{PA}(X_i^j)|} \to \mathbb{R}$ such that $X_i \in \mathbb{R}^{d+1}$ satisfies the following equation

$$X_i^j := f_i^j(X_i^{\mathrm{PA}(X_i^j)}, \varepsilon_i^j) \tag{1}$$

where $\mathrm{PA}(X_i^j) \subseteq \{X_i^1, \ldots, X_i^d\} \setminus \{X^j\}$ denotes the set of parent nodes of $X_i^j$ in the corresponding directed graph, and $(\varepsilon_i^j)_{j\in\{1,\ldots,d\}}$ are jointly independent noise variables with a joint distribution $\mathbb{P}_i^{\varepsilon}$ at time $i$. ♣

In [**CausalCP**], we focus our attention on linear sequential SCMs, where the function $f_i$ is linear in the parents and noise variables. A key assumption for a CCP to have its causal meaning under linear sequential SCMs is that there is no unobserved confounding between the covariates and the response. This is indeed a strong assumption, but we show that when unobserved confounding exists, CCPs may still be interpreted as changes in the observed causal mechanisms.

CCP detection can be formulated as testing a null hypothesis that there is no CCP in a time period. Intuitively, under the null hypothesis, no matter how we split the data into sub-intervals, we should be able to see that there is always an invariant set—a subset of the covariates such that the conditional distribution of $Y$ given this subset is the same across different sub-intervals.

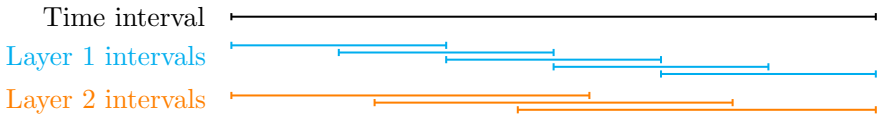



Figure 3: A simple illustration of seeded intervals. Each layer is assigned a different color for clarity. Intervals in each layer overlap and are of the same length.

CCP localization shares a similar intuition, to this end, two directions may be considered. First, as CCPs are a subset of RCPs, if RCPs are available, one can choose to detect the CCPs among them. The drawback of this approach is that, if RCPs need to be estimated, the RCP localization method needs to be powerful enough to capture all RCPs, and any estimation error would be passed on to the estimation error or CCPs.

Secondly, one can consider minimizing a suitable loss function. In [**CausalCP**], we propose a loss function called causal stability loss (CSL), such that if there is exactly one CCP in a time interval, the CCP is a minimizer of CSL. In order to localize multiple CCPs, we can combine CSL with existing multiple change point localization methods, such as the seeded binary segmentation algorithm [Kovács et al., 2023, SBS] with narrowest-over-threshold [Baranowski et al., 2019, NOT]. Compared to the standard top-down binary segmentation [Vostrikova, 1981], the advantage of SBS with NOT is that it is suitable for loss functions such as CSL whose minimizer may not be a CCP when there are multiple CCPs in an interval. The basic idea is to first generate "layers" of intervals, as illustrated in Figure 3, such that each layer contains overlapping intervals of a particular length. This way, one can ensure that each of the smallest intervals contains at most one change point. One can then estimate change points using a loss function starting from the shortest layer of intervals, and eliminate all longer intervals that contain the estimates estimated from the current layer.

**Data structure:** A sequence of $n$ independent observations $\{X_i, Y_i\}_{i=1}^n$ of a random variable $(X, Y)$ where $X \in \mathbb{R}^d$ are covariates and $Y \in \mathbb{R}$ is a univariate response.

**Causal question:** Are there changes in the causal relationship between $X$ and $Y$? If so, at which time points do the changes occur?

## 1.4 Causal inference with summary-level data

In the past decade, being able to work with summary-level data instead of individual-level data has become a very important feature of a statistical method in genetics research. On the one hand, due to privacy reasons, individual-level genetic datasets are not allowed to be made public without careful anonymization. On the other hand, there is a vast amount of summary-level data available from many international consortia such as biobanks [e.g., UK Biobank, 2024, Japan Biobank, 2024] and other sources [e.g., GWAS Catalog, 2024, FinnGen, 2024, All of Us, 2024].

Summary-level data commonly refers to the estimated marginal associations and standard errors between genetic variants and traits obtained from large-scale genome-wide association studies (GWAS). These summary-level data are also known as summary statistics. IV methods, referred to as Mendelian randomization (MR) in genetics research [see Sanderson et al., 2022, for an overview], can be adapted to work with summary statistics in the linear case. In fact, MR has become one of the most popular causal inference frameworks in genetic epidemiology, with using two-sample data (including two-sample summary statistics and two-sample individual-level data) being a more and more prominent data source in recent years (see Figure 4).

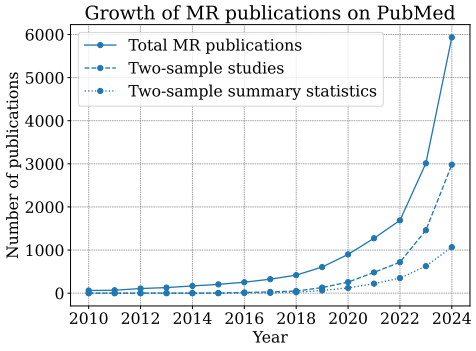



Figure 4: Total numbers of publications on PubMed containing the keyword "Mendelian randomization", along with counts of two-sample studies and those based on two-sample summary statistics, between 2010 and 2024. This figure is adapted from Hartwig et al. [2016, Figure 1] and updated with more recent data.

The two-sample setting here refers to when one has access to two separate datasets,


Figure 5: Illustration of IV settings. If the dashed arrow exists, the exclusion restriction criterion is violated.

often (assumed to be) collected from independent individuals, where one contains observations of the instruments $Z$ and the covariates $X$, and the other contains observations of the instruments $Z$ response $Y$. When only the summary statistics of these two datasets are available, this setting is referred to as the two-sample summary statistics setting. In [**SpaceTSIV**] (Chapter 4), we define (two-sample) summary statistics more formally based on the ordinary least squares (OLS) estimator, which is how summary statistics are normally computed in GWAS.

There are three classical assumptions of IV regularly mentioned in both economics and biomedical applications, namely, relevance, exchangeability, and the exclusion restriction criterion. Relevance relates to the strength of the instruments, and weak instruments are well-known to cause bias in IV methods such as two-stage least squares [Wooldridge, 2010]. This assumption is often tested in practice (where there are possibly multiple endogenous variables) by various under-identification and weak instrument tests such as the Cragg-Donald test [Cragg and Donald, 1993], Kleibergen-Paap test [Kleibergen and Paap, 2006], and conditional F test [Angrist and Pischke, 2009, Sanderson and Windmeijer, 2016]. On the contrary, the other two assumptions are in general untestable and often justified by domain knowledge. An equivalent way to express the exclusion restriction criterion is based on conditional independence: given instruments $Z$, covariates $X$, a response $Y$, and unobserved variables $H$, the following conditional independence holds if $Z$ are valid instruments

$$Z \perp\!\!\!\perp Y \mid X, H,$$

where we can also see that this condition cannot be tested as the variables $H$ are unobserved. Nevertheless, there are ways to falsify this assumption by testing its implications on the observables. For example, Sargan's test [Sargan, 1958] tries to falsify the validity of instruments in the over-identified setting, illustrated in the right plot of Figure 5. The intuition is the following. Consider the simplest case where there is a single covariate $X$, if the instruments are all valid, then the estimated causal effects based on each of these instruments should be similar; if an instrument is invalid, then the estimated causal effects based on those would appear to be "abnormal" compared to the others.

The majority of linear IV methods assume that there are at least as many instruments as the covariates in order to identify the causal effects. One exception was recently made by Pfister and Peters [2022], who show that under certain conditions, the causal parents (as well as their causal effects) are identifiable even when there are fewer instruments than covariates. The key assumption here is that the causal effects are sparse. In

[**SpaceTSIV**], we generalize this approach to the two-sample summary statistics setting.

We now discuss some specific features and challenges of MR applications that are beyond the scope of [**SpaceTSIV**], but could be of interest to investigate in the future. Firstly, the number of genetic variants is large, typically in millions. Genetic variants are also often highly correlated, known as linkage disequilibrium (LD), due to how we inherit our genes from our parents. This means selecting valid instruments among genetic variants can be very challenging [Garfield et al., 2023, Paz et al., 2023]. If genetic variants included in an MR analysis are highly correlated, it can lead to the causal effects being under- (or weakly-)identified, even though the number of instruments is larger than the number of covariates.

Secondly, when using two-sample summary statistics, it is often assumed that the two sets of individuals are similar. It is possible to address certain heterogeneity when the individual-level data is available [Zhao et al., 2019], but it has not been investigated given only summary statistics.

Thirdly, genetic variants that exist in only a very low percentage of the population, commonly referred to as "rare variants", may violate the positivity assumption, and whether rare variants are useful in MR is still under debate [Gibson, 2012, Zuk et al., 2014].

Lastly, the genetic variants are commonly coded as 0, 1, and 2, representing no mutation, mutation on one strand, and mutation on both strands in each location. This means that with linear MR, it is implicitly assumed that the incremental effect of the mutation is constant, which may be inadequate if the underlying relationship is non-linear.

**Data structure:** Individual-level data of a set of instrumental variables $Z \in \mathbb{R}^m$, a set of covariates $X \in \mathbb{R}^d$, and a univariate response $Y \in \mathbb{R}$ are inaccessible, but marginal OLS estimates and standard error of association between $Z$ and $X$ as well between $Z$ and $Y$ are. The correlation matrices of $Z$ and $X$ are also available.

**Causal question:** How to identify the direct causes of $Y$ among a set of (upstream) covariates $X$ and estimate their causal effects, where unobserved confounding exists between $X$ and $Y$, without access to individual level data?

# 2 Supervised learning and model analysis with compositional data

Shimeng Huang, Elisabeth Ailer, Niki Kilbertus, and Niklas Pfister

**Abstract**

Supervised learning, such as regression and classification, is an essential tool for analyzing modern high-throughput sequencing data, for example in microbiome research. However, due to the compositionality and sparsity, existing techniques are often inadequate. Either they rely on extensions of the linear log-contrast model (which adjust for compositionality but cannot account for complex signals or sparsity) or they are based on black-box machine learning methods (which may capture useful signals, but lack interpretability due to the compositionality). We propose `KernelBiome`, a kernel-based nonparametric regression and classification framework for compositional data. It is tailored to sparse compositional data and is able to incorporate prior knowledge, such as phylogenetic structure.

`KernelBiome` captures complex signals, including in the zero-structure, while automatically adapting model complexity. We demonstrate on par or improved predictive performance compared with state-of-the-art machine learning methods on 33 publicly available microbiome datasets. Additionally, our framework provides two key advantages: (i) We propose two novel quantities to interpret contributions of individual components and prove that they consistently estimate average perturbation effects of the conditional mean, extending the interpretability of linear log-contrast coefficients to nonparametric models. (ii) We show that the connection between kernels and distances aids interpretability and provides a data-driven embedding that can augment further analysis. `KernelBiome` is available as an open-source Python package on PyPI and at https://github.com/shimenghuang/KernelBiome.

## Author summary

In recent years, advances in gene sequencing technology have allowed scientists to examine entire microbial communities within genetic samples. These communities interact with their surroundings in complex ways, potentially benefiting or harming the host they

inhabit. However, analyzing the microbiome – the measured microbial community – is challenging due to the compositionality and sparsity of the data.

In this study, we developed a statistical framework called `KernelBiome` to model the relationship between the microbiome and a target of interest, such as the host's disease status. We utilized a type of machine learning model called kernel methods and adapted them to handle the compositional and sparse nature of the data, while also incorporating prior expert knowledge.

Additionally, we introduced two new measures to help interpret the contributions of individual compositional components. Our approach also demonstrated that kernel methods increase interpretability in analyzing microbiome data. To make `KernelBiome` as accessible as possible, we have created an easy-to-use software package for researchers and practitioners to apply in their work.

## 2.1 Introduction

Compositional data, that is, measurements of parts of a whole, are common in many scientific disciplines. For example, mineral compositions in geology [Buccianti et al., 2006], element concentrations in chemistry [Pesenson et al., 2015], species compositions in ecology [Jackson, 1997] and more recently high-throughput sequencing reads in microbiome science [Li, 2015].

Mathematically, any $p$-dimensional composition—by appropriate normalization—can be represented as a point on the simplex

$$\mathbb{S}^{p-1} \coloneqq \{x \in [0,1]^p \mid \textstyle\sum_{j=1}^{p} x^j = 1\}.$$

This complicates the statistical analysis, because the sum-to-one constraint of the simplex induces non-trivial dependencies between the components that may lead to false conclusions, if not appropriately taken into account.

The statistics community has developed a substantial collection of parametric analysis techniques to account for the simplex structure. The most basic is the family of Dirichlet distributions. However, as pointed out already by Aitchison [1982], Dirichlet distributions cannot capture non-trivial dependence structures between the composition components and are thus too restrictive. Aitchison [1982] therefore introduced the *log-ratio* approach. It generates a family of distributions by projecting multivariate normal distributions into $\mathbb{S}^{p-1}$ via an appropriate log-ratio transformation (e.g., the additive log-ratio, centered log-ratio [Aitchison, 1982], or isometric log-ratio [Egozcue et al., 2003]). The resulting family of distributions results in parametric models on the simplex that are rich enough to capture non-trivial dependencies between the components (i.e., beyond those induced by the sum-to-one constraint). The log-ratio approach has been extended and adapted to a range of statistical problems [e.g., Aitchison, 1985, Tsagris et al., 2011, Aitchison, 1983, Aitchison and Greenacre, 2002, Friedman and Alm, 2012].

For supervised learning tasks the log-ratio approach leads to the *log-contrast model* [Aitchison and Bacon-Shone, 1984]. An attractive property of the log-contrast model is that its coefficients quantify the effect of a multiplicative perturbation (i.e., fractionally

increasing one component while adjusting the others) on the response. While several extensions of the log-contrast model exist [e.g., Lin et al., 2014, Shi et al., 2016, Combettes and Müller, 2021, Simpson et al., 2021, Ailer et al., 2024], its parametric approach to supervised learning has two major shortcomings that become particularly severe when applied to high-dimensional and zero-inflated high-throughput sequencing data [Tsilimigras and Fodor, 2016, Gloor et al., 2017]. Firstly, since the logarithm is not defined at zero, the log-contrast model cannot be directly applied. A common fix is to add so-called pseudo-counts, a small non-zero constant, to all (zero) entries [Kaul et al., 2017, Lin and Peddada, 2020]. More sophisticated replacements exist as well [e.g., Martín-Fernández et al., 2003, Fernandes et al., 2013, De La Cruz and Kreft, 2018], however, they often rely on knowing the nature of the zeros (e.g., whether they are structural or random), which is typically not available in practice and difficult to estimate. In any case, the downstream analysis will strongly depend on the selected zero imputation scheme [Park et al., 2022]. Secondly, the relationships between individual components (e.g., species) and the response are generally complex. For example, in human microbiome settings, a health outcome may depend on interactions or on the presence or absence of species. Both cannot be captured by the linear structure of the log-contrast model.

We propose to solve the supervised learning task using a nonparametric kernel approach, which is able to handle complex signals and avoid arbitrary zero-imputation. To be of use in biological applications, there are two components to a supervised analysis: (i) estimating a predictive model that accurately captures signals in the data and (ii) extracting meaningful and interpretable information from the estimated model. For (i), it has been shown that modern machine learning methods are capable of creating highly predictive models by using microbiome data as covariates and phenotypes as responses [e.g., Pasolli et al., 2016, Knight et al., 2018, Zhou and Gallins, 2019, Cammarota et al., 2020]. In particular, several approaches have been proposed where kernels are used to incorporate prior information [Chen and Li, 2013, Randolph et al., 2018], as a way to utilize the compositional structure [Ramon et al., 2021, Di Marzio et al., 2015, Tsagris and Athineou, 2021] and to construct association tests [Zhao et al., 2015a, Wilson et al., 2021, Huang et al., 2022]. Our proposed framework extends these works by providing new post-analysis techniques (e.g., the compositional feature influence) that respect the compositional structure. Recently, Park et al. [2022], Li et al. [2023] used the radial transformation to argue that kernels on the sphere provide a natural way of analyzing compositions with zeros and similar to our work suggest using the kernel embeddings in a subsequent analysis. Part (ii) is related to the fields of explainable artificial intelligence[Samek et al., 2019] and interpretable machine learning [Molnar, 2020], which focus on extracting information from predictive models. These types of approaches have also received growing attention in the context of microbiome data [Topçuoğlu et al., 2020, Gou et al., 2021, Ruaud et al., 2022]. However, to the best of our knowledge, none of these procedures have been adjusted to account for the compositional structure. As we show in Section 2.2.1, not accounting for the compositionality may invalidate the results.

`KernelBiome`, see Fig 2.1.1, addresses both (i) by providing a regression and classification procedure based on kernels targeted to the simplex and (ii) by providing a principled way of analyzing the estimated models. Our contributions are fourfold: (1)

We develop a theoretical framework for using kernels on compositional data. While using kernels to analyze various aspects of compositional data is not a new idea, a comprehensive analysis and its connection to existing approaches has been missing. In this work, we provide a range of kernels that each capture different aspects of the simplex structure, many of which have not been previously applied to compositional data. For all kernels, we derive novel, positive-definite weighted versions that allow incorporating prior information between the components. Additionally, we show that the distance associated with each kernel can be used to define a kernel-based scalar summary statistic. (2) We propose a theoretically justified analysis of kernel-based models that accounts for compositionality. Firstly, we introduce two novel quantities for measuring the effects of individual features that explicitly take the compositionality into account and prove that these can be consistently estimated. Secondly, we build on known connections between kernels and distance measures to advocate for using the kernel embedding from the estimated model to create visualizations and perform follow-up distance-based analyses that respect the compositionality. (3) We draw connections between `KernelBiome` and log-contrast-based analysis techniques. More specifically, we connect the Aitchison kernel to the log-contrast model, prove that the proposed compositional feature influence in this case reduces to the log-contrast coefficients, and show that our proposed weighted Aitchison kernel is related to the recently proposed tree-aggregation method of log-contrast coefficients [Bien et al., 2021]. Importantly, these connections ensure that `KernelBiome` reduces to standard log-contrast analysis techniques whenever simple log-contrast models are capable of capturing most of the signal. This is also illustrated by our experimental results. (4) We propose a data-adaptive selection framework that allows to compare different kernels in a principled fashion.

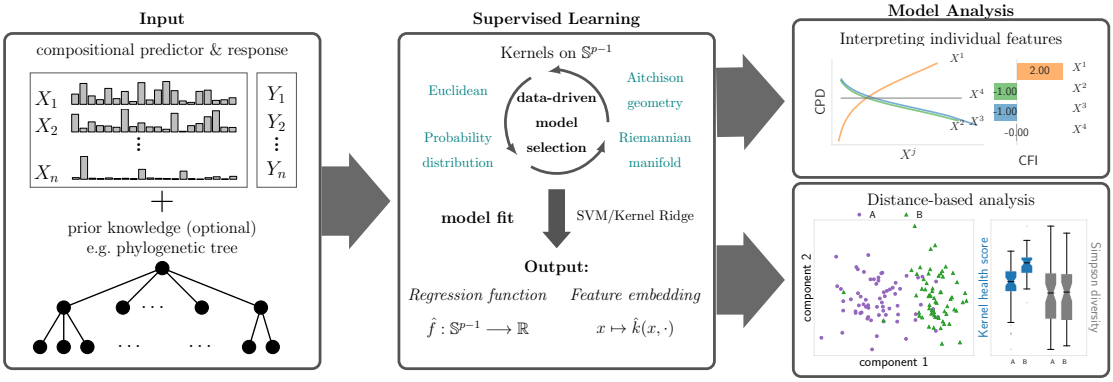



Figure 2.1.1: Overview of `KernelBiome`. We start from a paired dataset with a compositional predictor $X$ and a response $Y$ and optional prior knowledge on the relation between components in the compositions (e.g., via a phylogenetic tree). We then select a model among a large class of kernels which best fits the data. This results in an estimated model $\hat{f}$ and embedding $\hat{k}$. Finally, these can be analyzed while accounting for the compositional structure.

The paper is structured as follows. In Section 2.2, we introduce the supervised learning

task, define two quantities for analyzing individual components (Section 2.2.1), give a short introduction to kernel methods and how to apply our methodology (Section 2.2.2), and present the full `KernelBiome` framework (Section 2.2.3). Finally, we illustrate the advantages of `KernelBiome` in the experiments in Section 2.3.

## 2.2 Methods

We consider the setting in which we observe $n$ independent and identically distributed (i.i.d.) observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a random variable $(X, Y)$ with $X \in \mathbb{S}^{p-1}$ a compositional predictor and $Y \in \mathbb{R}$ a real-valued response variable (by which we include categorical responses). Supervised learning attempts to learn a relationship between the response $Y$ and the dependent predictors $X$. In this work, we focus on conditional mean relationships. More specifically, we are interested in estimating the conditional mean of $Y$, that is, the function

$$f^* : x \mapsto \mathbb{E}[Y \mid X = x]. \tag{2.2.1}$$

We assume that $f^* \in \mathcal{F} \subseteq \{f \mid f : \mathbb{S}^{p-1} \to \mathbb{R}\}$, where $\mathcal{F}$ is a function class determined by the regression (or classification) procedure.

While estimating and analyzing the conditional mean is well established for predictors in Euclidean space, there are two factors that complicate the analysis when the predictors are compositional. (i) While it is possible to directly apply most standard regression procedures designed for $X \in \mathbb{R}^p$ also for $X \in \mathbb{S}^{p-1}$, it turns out that many approaches are ill-suited to approximate functions on the simplex. (ii) Even if one accurately estimates the conditional mean function $f^*$, the simplex constraint complicates any direct assessment of the influence and importance of individual components of the compositional predictor. In this work, we address both issues and propose a nonparametric framework for regression and classification analysis for compositional data.

### 2.2.1 Interpreting individual features

Our goal when estimating the conditional mean $f^*$ given in (2.2.1) is to gain insight into the relationship between the response $Y$ and predictors $X$. For example, when fitting a log-contrast model (see Example 2.2.2), the estimated coefficients provide a useful tool to generate hypotheses about which features affect the response and thereby inform follow-up experiments. For more complex models, such as the nonparametric methods proposed in this work, direct interpretation of a fitted model $\hat{f}$ is difficult. Two widely applicable measures due to Friedman [2001] are the following: (i) Relative influence, which assigns each coordinate $j$ a scalar influence value given by the expected partial derivative $\mathbb{E}[\frac{d}{dx^j}\hat{f}(X)]$ and (ii) partial dependence plots, which are constructed by plotting, for each coordinate $j$, the function $z \mapsto \mathbb{E}[\hat{f}(X^1, \ldots, X^{j-1}, z, X^{j+1}, \ldots, X^p)]$. However, directly applying these measures on the simplex is not possible as we illustrate in Fig 2.D.2 in Appendix 2.D. The intuition is that both measures evaluate the function $\hat{f}$ outside the simplex. An adaptation of the relative influence (or elasticity in the econometrics literature) to compositions based on the Aitchison geometry has recently

been proposed by Morais and Thomas-Agnan [2021]. We adapt the relative influence without relying on the log-ratio transform and hence allow for more general function classes.

Our approach is based on two coordinate-wise perturbations on the simplex. For any $j \in \{1, \dots, p\}$ and $x \in \mathbb{S}^{p-1}$, define (i) for $c \in [0, \infty)$ the function $\psi_j(x, c) \in \mathbb{S}^{p-1}$ to be the composition resulting from multiplying the $j$-th component by $c$ and then scaling the entire vector back into the simplex, and (ii) for $c \in [0, 1]$ the function $\varphi_j(x, c) \in \mathbb{S}^{p-1}$ to be the composition that consists of fixing the $j$-th coordinate to $c$ and then rescaling all remaining coordinates such that the resulting vector lies in the simplex. Each perturbation can be seen as a different way of intervening on a single coordinate while preserving the simplex structure. More details are given in Appendix 2.A. Based on the first perturbation, we define the *compositional feature influence* (CFI) of component $j \in \{1, \dots, p\}$ for any differentiable function $f : \mathbb{S}^{p-1} \to \mathbb{R}$ by

$$\text{(CFI)} \qquad I_f^j := \mathbb{E}\Big[\tfrac{d}{dc} f(\psi_j(X, c))\big|_{c=1}\Big]. \tag{2.2.2}$$

Similarly, we adapt partial dependence plots using the second perturbation. Define the *compositional feature dependence* (CPD) of component $j \in \{1, \dots, p\}$ for any function $f : \mathbb{S}^{p-1} \to \mathbb{R}$ by

$$\text{(CPD)} \qquad S_f^j : z \mapsto \mathbb{E}[f(\varphi_j(X, z))] - \mathbb{E}[f(X)]. \tag{2.2.3}$$

In practice, we can compute Monte Carlo estimates of both quantities by replacing expectations with empirical means. We denote the corresponding estimators by $\hat{I}_f^j$ and $\hat{S}_f^j$, respectively (see Appendix 2.A for details).

The following proposition connects the CFI and CPD to the coefficients in a log-contrast function.

**Proposition 2.2.1** (CFI and CPD in the log-contrast model)**.** *Let $f : x \mapsto \beta^T \log(x)$ with $\sum_{j=1}^p \beta_j = 0$, then the CFI and CPD are given by*

$$I_f^j = \beta_j \quad \textit{and} \quad S_f^j : z \mapsto \beta_j \log\left(\tfrac{z^j}{1-z^j}\right) + c,$$

*respectively, where $c \in \mathbb{R}$ is a constant depending on the distribution of $X$ but not on $z$ and satisfies $c = 0$ if $\beta_j = 0$.*

A proof is given in Appendix 2.F. The proposition shows that the CFI and CPD are generalizations of the $\beta$-coefficients in the log-contrast model. The following example provides further intuition.

**Example 2.2.2** (CFI and CPD in a log-contrast model)**.** Consider a log-contrast model $Y = f(X) + \epsilon$ with $f : x \mapsto 2\log(x^1) - \log(x^2) - \log(x^3)$.

The CFI and CPD for the true function $f$ — estimated based on $n = 100$ i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n)$ with $X_i$ compositional log-normal — are shown in Fig 2.2.2. ♠

The following theorem highlights the usefulness of the CFI and CPD by establishing when they can be consistently estimated from data.

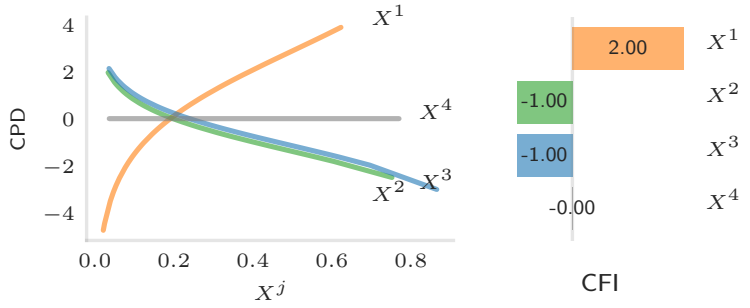



Figure 2.2.2: Visualization of the CPD (left) and CFI (right) based on $n = 100$ samples and the true function $f$. Since $\beta_4 = 0$ in this example the 4-th component has no effect on the value of $f$ resulting in a CFI of zero and a flat CPD. Since we are not estimating $f$, the CFI values exactly correspond to the $\beta$-coefficients in this example.

**Theorem 2.2.3** (Consistency). *Assume $\hat{f}_n$ is an estimator of the conditional mean $f^*$ given in (2.2.1) based on $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d..*

*(i) If $\frac{1}{n}\sum_{i=1}^n \left\|\nabla \hat{f}_n(X_i) - \nabla f^*(X_i)\right\|_2 \xrightarrow{P} 0$ as $n \to \infty$ and $\mathbb{E}[(\nabla f^*(X_i))^2] < \infty$, then it holds for all $j \in \{1, \dots, p\}$ that*

$$\hat{I}^j_{\hat{f}_n} \xrightarrow{P} I^j_{f^*} \quad \text{as } n \to \infty.$$

*(ii) If $\sup_{x \in \text{supp}(X)} |\hat{f}_n(x) - f^*(x)| \xrightarrow{P} 0$ as $n \to \infty$ and $\text{supp}(X) = \{w/(\sum_j w^j) \mid w \in \text{supp}(X^1) \times \cdots \times \text{supp}(X^p)\}$, then it holds for all $j \in \{1, \dots, p\}$ and all $z \in [0, 1]$ with $z/(1-z) \in \text{supp}(X^j / \sum_{\ell \neq j} X^\ell)$ that*

$$\hat{S}^j_{\hat{f}_n}(z) \xrightarrow{P} S^j_{f^*}(z) \quad \text{as } n \to \infty.$$

A proof is given in Appendix 2.F.1 and the result is demonstrated on simulated data in Fig 2.D.1 in Appendix 2.D. The theorem shows that the CFI is consistently estimated as long as the derivative of $f^*$ is consistently estimated, which can be ensured for example for the kernel methods discussed in Section 2.2.2. In contrast, the CPD only requires the function $f^*$ itself to be consistently estimated. The additional assumption on the support ensures that the perturbation $\varphi_j$ used in the CPD remains within the support. If this assumption is not satisfied one needs to ensure that the estimated function extrapolates beyond the sample support. Interpreting the CPD therefore requires caution.

### 2.2.2 Kernel methods for compositional data analysis

Before presenting our proposed weighted and unweighted kernels, we briefly review the necessary background on kernels and their connection to distances. Kernel methods are

a powerful class of nonparametric statistical methods that are particularly useful for data from non-standard (i.e., non-Euclidean) domains $\mathcal{X}$. The starting point is a symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called kernel. Kernels encode similarities between points in $\mathcal{X}$, i.e., large values of $k$ correspond to points that are similar and small values to points that are less similar. Instead of directly analyzing the data on $\mathcal{X}$, kernel methods map it into a well-behaved feature space $\mathcal{H}_k \subseteq \{f \mid f : \mathcal{X} \to \mathbb{R}\}$ called reproducing kernel Hilbert space (RKHS), whose inner product preserves the kernel induced similarity.

Here, we consider kernels on the simplex, that is, $\mathcal{X} = \mathbb{S}^{p-1}$. The conditional mean function $f^*$ given in (2.2.1) can then be estimated by optimizing a loss over $\mathcal{H}_k$, for an appropriate kernel $k$ for which $\mathcal{H}_k$ is sufficiently rich, i.e., $f^* \in \mathcal{H}_k$. The representer theorem [e.g., Schölkopf et al., 2002] states that such an optimization over $\mathcal{H}_k$ can be performed efficiently. Formally, it states that the minimizer of an arbitrary convex loss function $L : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ of the form

$$\hat{f} = \underset{f \in \mathcal{H}_k}{\arg\min}\, L\big((Y_1, f(X_1)), \ldots, (Y_n, f(X_n))\big) + \lambda ||f||^2_{\mathcal{H}_k},$$

with $\lambda > 0$ a penalty parameter, has the form $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(X_i, \cdot)$ for some $\hat{\alpha} \in \mathbb{R}^n$. This means that instead of optimizing over a potentially infinite-dimensional space $\mathcal{H}_k$, it is sufficient to optimize over the $n$-dimensional parameter $\hat{\alpha}$. Depending on the loss function, this allows to construct efficient regression and classification procedures, such as kernel ridge regression and support vector machines [e.g., Schölkopf et al., 2002].

The performance of the resulting prediction model depends on the choice of kernel as this determines the function space $\mathcal{H}_k$. A useful way of thinking about kernels is via their connection to distances. In short, any kernel $k$ induces a unique semi-metric $d_k$ and vice versa. More details are given in Appendix 2.E. This connection has two important implications. Firstly, it provides a natural way for constructing kernels based on established distances on the simplex. The intuition being that a distance, which is large for observations with vastly different responses and small otherwise, leads to an informative feature space $\mathcal{H}_k$. Secondly, it motivates using the kernel-induced distance, see Section 2.2.3.2.

#### 2.2.2.1 Kernels on the simplex

We consider four types of kernels on the simplex, each related to different types of distances. A full list with all kernels and induced distances is provided in Appendix 2.G. While most kernels have previously appeared in the literature, we have adapted many of the kernels to fit into the framework provided here, e.g., added zero-imputation for Aitchison kernels and updated the parametrization for the probability distribution kernels.

**Euclidean:** These are kernels that are constructed by restricting kernels on $\mathbb{R}^p$ to the simplex. Any such restriction immediately guarantees that the restricted kernel is again a kernel. However, the induced distances are not targeted to the simplex and therefore can be unnatural choices. In `KernelBiome`, we have included the linear kernel and the

radial basis function (RBF) kernel. The RBF kernel is $L^p$-universal [e.g., Sriperumbudur et al., 2011] which means that it can approximate any integrable function (in the large sample limit). However, this does not necessarily imply good performance for finite sample sizes.

**Aitchison geometry:** One way of incorporating the simplex structure is to use the Aitchison geometry. Essentially, this corresponds to mapping points from the interior of the simplex via the centered log-ratio transform into $\mathbb{R}^p$ and then using the Euclidean geometry. This results in the Aitchison kernel for which the induced RKHS is equal to the log-contrast functions. In particular, applying kernel ridge regression with an Aitchison kernel corresponds to fitting a log-contrast model with a penalty on the coefficients. As the centered log-ratio transform is only defined for interior points in the simplex, we add a hyperparameter to the kernels that shift them away from zero. From this perspective, the commonly added pseudo-count constant added to all components becomes a tuneable hyperparameter of our method, rather than a fixed ad-hoc choice during data pre-processing. Thereby, our modified Aitchison kernel respects the fact that current approaches to zero-replacement or imputation are often not biologically justified, yet may impact predictive performance. Our proposed zero-imputed Aitchison kernel comes with two advantages over standard log-contrast modelling: (1) A principled adjustment for zeros and (2) an efficient form of high-dimensional regularization that performs well across a large range of our experiments. In `KernelBiome`, we include the Aitchison kernel and the Aitchison-RBF kernel which combines the Aitchison and RBF kernels.

**Probability distributions:** Another approach to incorporate the simplex structure into the kernel is to view points in the simplex as discrete probability distributions. This allows us to make use of the extensive literature on distances between probability distributions to construct kernels. In `KernelBiome`, we have adapted two classes of such kernels: (1) A parametric class based on generalized Jensen-Shannon distances due to Topsøe [2003], which we call generalized-JS, and (2) a parametric class based on the work by Hein and Bousquet [2005], which we call Hilbertian. Together they contain many well-established distances such as the total variation, Hellinger, Chi-squared, and Jensen-Shannon distance. All resulting kernels allow for zeros in the components of compositions.

**Riemannian manifold:** Finally, the simplex structure can be incorporated by using a multinomial distribution which has a parameter in the simplex. Lafferty et al. [2005] show that the geometry of multinomial statistical models can be exploited by using kernels based on the heat equation on a Riemannian manifold. The resulting kernel is known as the heat-diffusion kernel and has been observed to work well with sparse data.

#### 2.2.2.2 Including prior information into kernels

All kernels introduced in the previous section (and described in detail in Appendix 2.G.1 are invariant under permutations of the compositional components. They therefore do not take into account any relation between the components. In many applications, one may however have prior knowledge about the relation between the components. For ex-

ample, if the compositional predictor consists of relative abundances of microbial species, information about the genetic relation between different species encoded in a phylogenetic tree may be available. Therefore, we provide the following way to incorporate such relations. Assume the prior information has been expressed as a positive semi-definite weight matrix $W \in \mathbb{R}^{p\times p}$ with non-negative entries (e.g., using the UniFrac-Distance [Lozupone and Knight, 2005] as shown in Appendix 2.B.3, where the $ij$-th entry corresponds to the strength of the relation between components $i$ and $j$. We can then incorporate $W$ directly into our kernels. To see how this works, consider the special case where the kernel $k$ can be written as $k(x,y) = \sum_{i=1}^{p} k_0(x^i, y^i)$ for a positive definite kernel $k_0 : [0,1] \times [0,1] \to \mathbb{R}$. Then, the weighted kernel

$$k_W(x,y) \coloneqq \sum_{i,j=1}^{p} W_{i,j} \cdot k_0(x^i, y^j) \tag{2.2.4}$$

is positive definite and incorporates the prior information in a natural way. If two components $i$ and $j$ are known to be related (corresponding to large values of $W_{i,j}$), the kernel $k_W$ takes the similarity across these components into account. In Appendix 2.B.2, we show that the probability distribution kernels and the linear kernel can be expressed in this way and propose similar weighted versions for the remaining kernels.

An advantage of our framework is that it defaults to the log-contrast model when more complex models fail to improve the prediction (due to the zero-shift in our proposed Aitchison kernel and the kernel-based regularization, this correspondence is however not exact). We now show that for the weighted Aitchison kernel, the RKHS consists of log-contrast functions with equal coefficients across the weighted blocks, this is similar to how Bien et al. [2021] incorporate prior information into log-contrast models.

**Proposition 2.2.4** (weighted Aitchison kernel RKHS)**.** *Let $P_1, \dots, P_m \subseteq \{1, \dots, p\}$ be a disjoint partition and $W \in \mathbb{R}^{p\times p}$ the weight matrix defined for all $i, j \in \{1, \dots, p\}$ by $W_{i,j} \coloneqq \sum_{\ell=1}^{m} \frac{1}{|P_\ell|} \mathbb{1}_{\{i,j\in P_\ell\}}$. Let $k_W$ be the weighted Aitchison kernel given in Appendix 2.G.2 (but without zero imputation and on the open simplex). Then, it holds that*

$$f \in \mathcal{H}_{k_W} \quad \Leftrightarrow \quad f = \beta^\top \log(\cdot)$$

*for some $\beta \in \mathbb{R}^p$ satisfying (1) $\sum_{j=1}^{p} \beta_j = 0$ and (2) for all $\ell \in \{1, \dots, m\}$ and $i, j \in P_\ell$ it holds $\beta_i = \beta_j$*

A proof is given in Appendix 2.F.2. Combined with Proposition 2.2.1, this implies that the CFI values are equal across the equally weighted blocks $P_1, \dots, P_m$, which is demonstrated empirically in Section 2.3.2 and Section 2.3.4.

### 2.2.3 KernelBiome framework

For a given i.i.d. dataset $(X_1, Y_1), \dots, (X_n, Y_n)$, `KernelBiome` first runs a data-driven model selection, resulting in an estimated regression function $\hat{f}$ and a specific kernel $\hat{k}$ (see Fig 2.1.1). Then, the feature influence properties (CFI, CPD) and embedding induced by $\hat{k}$ are analyzed in a way that respects compositionality.

#### 2.2.3.1 Model selection

We propose the following two step data-driven selection procedure.

1. Select the best kernel $\hat{k}$ with the following hierarchical CV:
   - Fix a kernel $\tilde{k}$, i.e., a type of kernel and its kernel parameters.
   - Split the sample into $N_{\text{out}}$ random (or stratified) folds.
   - For each fold, use all other folds to perform a $N_{\text{in}}$-fold CV to select the best hyperparameter $\tilde{\lambda}$ and compute a CV score based on $\tilde{k}$ and $\tilde{\lambda}$ on the left-out fold.
   - Select the kernel $\hat{k}$ with the best average CV score.
2. Select the best hyperparameter $\hat{\lambda}$ for $\hat{k}$ using a $N_{\text{in}}$-fold CV on the full data. The final estimator $\hat{f}$ is then given by the kernel predictor based on $\hat{k}$ and $\hat{\lambda}$.

This CV scheme ensures that all parameters are selected in a data adaptive way. Up to a point, including more parameter settings into the CV makes the method more robust at the cost of additional run time. We provide sensible default choices for all parameters (see e.g., Table 2.B.1 in Appendix 2.B.1 for the default kernels), allowing practitioners to directly apply the method. In the `KernelBiome` implementation, the parameter grids for the kernel parameters and hyperparameters, as well as parameters of the CV including the type of CV, number of CV folds, and scoring can also be be adjusted manually, for example to reduce the run time.

#### 2.2.3.2 Model analysis

Firstly, as discussed in Section 2.2.1, we propose to analyze the fitted model $\hat{f}$ with the CPD and CFI. Other methods developed for functions on $\mathbb{R}^p$ do not account for compositionality and can be misleading. Secondly, the kernel embedding $\hat{k}$ can be used for the following two types of analyses.

**Distance-based analysis:** A key advantage of using kernels is that the fitted kernel $\hat{k}$ is itself helpful in the analysis. As discussed in Section 2.2.2, $\hat{k}$ induces a distance on the simplex that is well-suited to separate observations with different response values. We therefore suggest to utilize this distance to investigate the data further. Essentially, any statistical method based on distances can be applied. We specifically suggest using kernel PCA to project the samples into a two-dimensional space. As we illustrate in Section 2.3.3, such a projection can be used to detect specific groups or outliers in the samples and can also help understand how the predictors are used by the prediction model $\hat{f}$. As we are working with compositional data we need to be careful when looking at how individual components contribute to each principle component. Fortunately, the perturbation $\psi$ defined in Section 2.2.1 can again be used to construct informative component-wise measures. All details on kernel PCA and how to compute component-wise contributions for each principle component are provided in Appendix 2.E.2.

**Data-driven scalar summary statistics:** Practitioners often work with scalar summaries of the data as these are easy to communicate. A commonly used summary statistic

in ecology is $\alpha$-diversity which measures the variation within a community. The connection between kernels and distances provides a useful tool to construct informative scalar summary statistics by considering distances to a reference point $u$ in the simplex. Formally, for a fixed reference point $u \in \mathbb{S}^{p-1}$ define for all $x \in \mathbb{S}^{p-1}$ a corresponding closeness measure by $D^k(x) \coloneqq -d_k^2(x, u)$, where $d_k$ is the distance induced by the kernel $k$. This provides an easily interpretable scalar quantity. For example, if $Y$ is a binary indicator taking values *healthy* and *sick*, we could select $u$ to be the geometric median (the observation that has the smallest total distance to all the other observations based on the pairwise kernel distance) of all $X_i$ with $Y_i = \textit{healthy}$. Then, $D^k$ corresponds to a very simple health score (see Section 2.3.3 for a concrete example). A further example is given by selecting $u = (1/p, \ldots, 1/p)$ and considering points on the simplex as communities. Then, $u$ can be interpreted as the most diverse point in the simplex and $D^k$ corresponds to a data-adaptive $\alpha$-diversity measure. While such a definition of diversity does not necessarily satisfy all desirable properties for diversities [see e.g., Leinster and Cobbold, 2012], it is (1) symmetric with respect to switching of coordinates, (2) has an intuitive interpretation and (3) is well-behaved when combined with weighted kernels. Connections to established diversities also exist, for example, the linear kernel corresponds to a shifted version of the Gini-Simpson diversity (i.e., Gini-Simpson$(x) \coloneqq 1 - \sum_{j=1}^{p}(x^j)^2 = D^k(x) + \frac{p-2}{p}$).

#### 2.2.3.3 Run time complexity

The run time complexity of `KernelBiome` depends on the number of kernels $K$, the size of the hyperparameter grid $H$, and the number of inner CV folds $N_{\text{in}}$ and outer CV folds $N_{\text{out}}$. Since the run time complexity of kernel ridge regression and support vector machines is $O(n^3)$ (based on a straightforward implementation, actual implementation in available software libraries can achieve a more optimized run time), the total run time complexity of `KernelBiome` is $O(KHN_{\text{in}}N_{\text{out}}n^3)$. For example, the default parameter settings use 55 kernels, with 5-fold inner CV and 10-fold outer CV with a hyperparameter grid of size 10, resulting in 27,500 model fits, each of complexity $O(n^3)$. To reduce the run time we recommend reducing the number of kernels $K$, this can be particularly useful for prototyping. However, if possible, we recommend using the full list of kernels for a final analysis to avoid a decrease in predictive performance.

#### 2.2.3.4 Implementation

`KernelBiome` is implemented as a Python [van Rossum and Drake, 2009] package that takes advantage of the high-performance `JAX` [Bradbury et al., 2018] and `scikit-learn` [Pedregosa et al., 2011] libraries. All kernels introduced are implemented with `JAX`'s just-in-time compilation and automatically leverage accelerators such as GPU and TPU whenever available. `KernelBiome` provides fast computation of all kernels and distance metrics as well as easy-to-use procedures for model selection and comparison and procedures to estimate CPD and CFI, compute kernel PCA, and estimate scalar summary statistics. An illustration script for the package's usage can be found in the package

repository.

## 2.3 Results

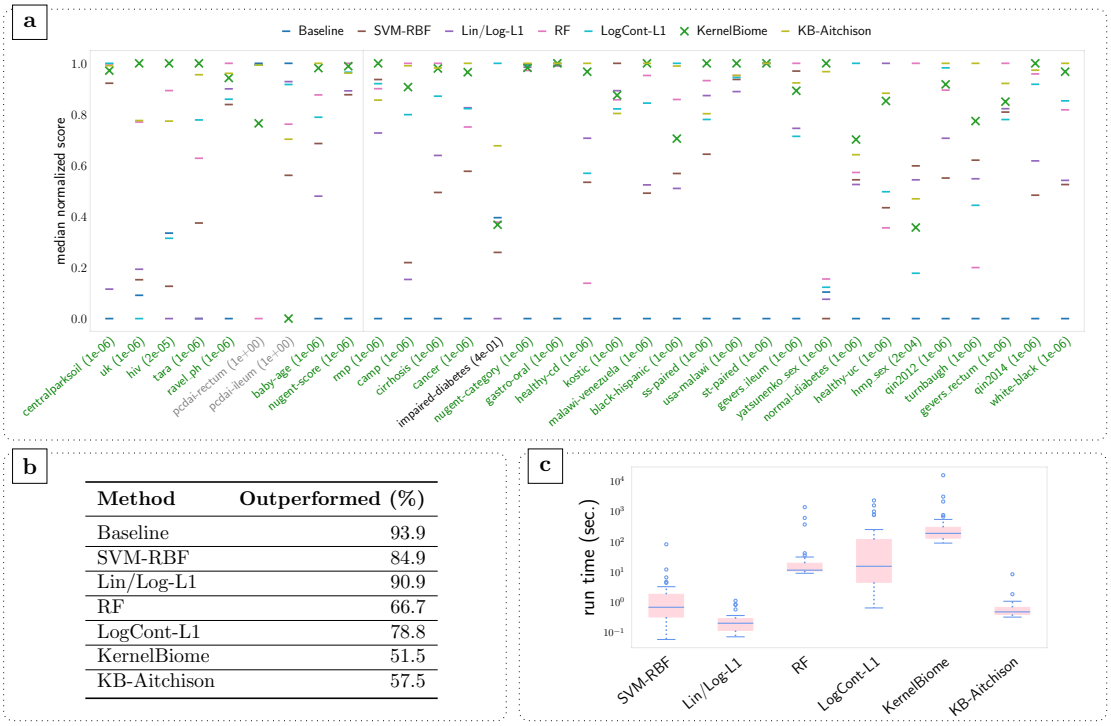
We evaluated `KernelBiome` on a total of 33 microbiome datasets. All datasets have been previously published and final datasets used in our experiments can be fully reproduced following the description in the GitHub repo https://github.com/shimenghuang/KernelBiome. A summary of the datasets including on the pre-processing steps, prediction task and references is provided in Table 2.C.1 in Appendix 2.C. First, in Section 2.3.1, we show that `KernelBiome` performs on par or better than existing supervised learning procedures for compositional data, while reducing to a powerfully regularized version of the log-contrast model if the prediction task is simple. In Section 2.3.2, we show on a semi-artificial dataset that including prior information can either improve or harm the predictive performance depending on whether or not it is relevant for the prediction. In Section 2.3.3, we illustrate the advantages of a full analysis with `KernelBiome`. Finally, in Section 2.3.4, we demonstrate how `KernelBiome` can incorporate prior knowledge, while preserving a theoretically justified interpretation.

### 2.3.1 State-of-the-art prediction performance

We compare the predictive performance of `KernelBiome` on all datasets with the following competitors: (i) `Baseline`, a naive predictor that uses the training majority class for classification and the training mean for regression as its prediction, (ii) `SVM-RBF`, a support vector machine with the RBF kernel, (iii) `Lin/Log-L1`, a linear/logistic regression with $\ell^1$-penalty (iv) `LogCont-L1`, a log-contrast regression with $\ell^1$ penalty with a half of the minimum non-zero relative abundance added as pseudo-count to remove zeros, and (v) `RF`, a random forest with 500 trees. For `SVM-RBF`, `Lin/Log-L1` and `RF` we use the `scikit-learn` implementations [Pedregosa et al., 2011] and choose the hyperparameters (bandwidth, max depth and all penalty parameters) based on a 5-fold CV. For `LogCont-L1`, we use the `c-lasso` package [Simpson et al., 2021] and the default CV scheme to chose the penalty parameter. We apply two versions of `KernelBiome`: (1) The standard version that adaptively chooses the kernel using $N_{\text{in}} = 5$, $N_{\text{out}} = 10$ (denoted `KernelBiome`), and (2) a version with fixed Aitchison kernel with $c$ equal to half of the minimum non-zero relative abundance (denoted `KB-Aitchison`). Both methods use a default hyperparameter grid of size 40, and we use kernel ridge regression as the estimator. We compared with using support vector regression instead of kernel ridge regression and the results are similar.

For the comparison we perform 20 random (stratified) 10-fold train/test splits and record the predictive performance (balanced accuracy for classification and root-mean-squared error (RMSE) for regression) on each test set. Fig 2.3.3 contains the summary results for the 33 datasets. Fig 2.3.3(a) gives an overview of the predictive performance. To make the comparison easier, the median test scores are normalized to between 0 and 1 based on the minimum and maximum scores of each dataset. More details of the predictive performance results including boxplots of scores for all tasks and precision-recall

curves for all classification tasks are provided in Figs 2.C.2 and 2.C.3 in Appendix 2.C. Moreover, we perform a Wilcoxon signed-rank test [Wilcoxon, 1992] on the test scores and the percentage of times a method is significantly outperformed by another is given in Fig 2.3.3(b). Lastly, run times of each method on the 33 datasets are shown in Fig 2.3.3(c).




| Method | Outperformed (%) |
|---|---|
| Baseline | 93.9 |
| SVM-RBF | 84.9 |
| Lin/Log-L1 | 90.9 |
| RF | 66.7 |
| LogCont-L1 | 78.8 |
| KernelBiome | 51.5 |
| KB-Aitchison | 57.5 |

Figure 2.3.3: (a) Comparison of predictive performance on 33 public datasets (9 regression and 24 classification tasks, separated by the grey vertical line in the figure) based on 20 random 10-fold CV. On the two datasets with grey tick labels no method significantly outperforms the baseline based on the Wilcoxon signed-rank test, meaning that there is little signal in the data. The ones in green are the datasets where `KernelBiome` significantly outperforms the baseline, while it does not on the single dataset with the black label. The corresponding p-values are provided in brackets. (b) Percentage of time a method is significantly outperformed by another based on the Wilcoxon signed-rank test. (c) Average run time of each method on each dataset. A significance level of 0.05 is used.

On all datasets `KernelBiome` achieves the best or close to best performance and (almost) always captures useful information (green labels in Fig 2.3.3(a)), indicating that the proposed procedure is well-adapted to microbiome data. The kernel which was selected most often by `KernelBiome` and the frequency with which it was selected are shown in Table 2.C.2 in Appendix 2.C. There are several interesting observations: (1)

Even though `KernelBiome` selects mostly the Aitchison kernel on rmp, it outperforms `KB-Aitchison`, we attribute this to the advantage of the data-driven zero-imputation. (2) On datasets were the top kernel is selected consistently (e.g., uk, hiv and tara) `KernelBiome` generally performs very well and in these cases strongly outperformed both log-contrast based methods `KB-Aitchison` and `LogCont-L1`. (3) The predictive performance is substantially different between `KB-Aitchison` and `LogCont-L1` which we see as an indication that the type of regularization (kernel-based vs $\ell^1$-regularization, respectively) is crucial in microbiome datasets.

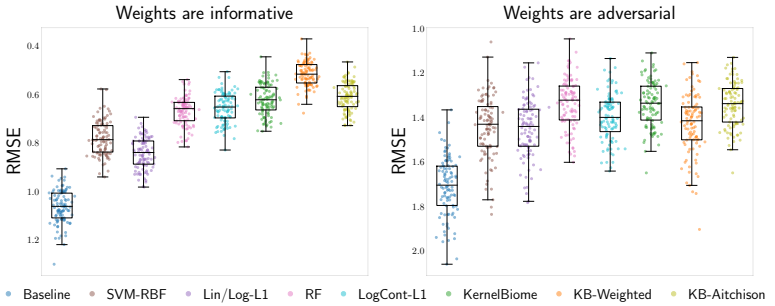
### 2.3.2 Predictive performance given prior information

In many applications, in particular in biology, prior information about a system is available and should be incorporated into the data analysis. As we show in Section 2.2.2.2, `KernelBiome` allows for incorporating prior knowledge on the relation between individual components (e.g., taxa). We will illustrate in this section that given informative prior knowledge, the predictive performance of `KernelBiome` can be improved, while if the prior knowledge is uninformative or incorrect, the predictive performance can be harmed. We conduct a semi-synthetic experiment based on the uk dataset. The dataset has 327 species and $n = 882$ samples. We generate the response $Y$ based on two processes: (DGP1) a linear log-contrast model where species under phylum Bacteroidetes all have coefficient $\beta_B$, species under phylum Proteobacteria all have coefficient $\beta_P$, and all coefficients corresponding to other species are set to 0; (DGP2) a linear log-contrast model where the first half of the species under Bacteroidetes are given coefficient $-\beta_B$ while the second half are given coefficient $\beta_B$, similarly for Proteobacteria.

We construct a weight matrix based on the phylum each species belongs to (similar as in Prop. 2.2.4). By construction these weights are informative if the data are generated from DGP1, but not if the data are generated from DGP2. For each DGP, we sample 100 data points for training and another 100 data points for testing, repeated for 200 times. We compare the predictive performance of all methods mentioned in Section 2.3.1 and weighted `KernelBiome` (KB-Weighted) in Fig 2.3.4. The results show that when the weights are indeed informative, KB-Weighted achieves the best performance and is significantly better than the unweighted version (p-value $2.34 \times 10^{-18}$ based on Wilcoxon signed-rank test). On the other hand, when the weights do not align with the underlying generating mechanism, the predictive performance of KB-Weighted can be significantly worse than the unweighted one (p-value $1.25 \times 10^{-9}$ based on Wilcoxon signed-rank test). A table containing the number of other methods a method is significantly outperformed by under the two DGPs is also provided in Fig 2.3.4. All methods significantly outperform the baseline in this example, and the corresponding p-values are all below $2 \times 10^{-17}$.

### 2.3.3 Model analysis with KernelBiome

As shown in the previous section, `KernelBiome` results in fitted models with state-of-the-art prediction performance. This is useful because supervised learning procedures




| Method | DGP1 | DGP2 |
|---|---|---|
| Baseline | 7 | 7 |
| SVM-RBF | 5 | 4 |
| Lin/Log-L1 | 6 | 5 |
| RF | 4 | 0 |
| LogCont-L1 | 3 | 3 |
| KernelBiome | 2 | 0 |
| KB-Weighted | 0 | 4 |
| KB-Aitchison | 1 | 0 |

Figure 2.3.4: Left and middle: predictive performance of weighted `KernelBiome` when the given weights are informative (DGP1) and adversarial (DGP2) based on 200 repetitions. Right: the number of other methods each method is significantly outperformed by based on Wilcoxon signed-rank test (siginificance level 0.05), under DGP1 and DGP2.

can be used in two types of applications: (1) To learn a prediction model that has a direct application, e.g., as a diagnostic tool, or (2) to learn a predictive model as an intermediate step of an exploratory analysis to find out what factors could be driving the response. As discussed above (2) requires us to take the compositional nature of the predictors into account to avoid misleading conclusions. We show how the `KernelBiome` framework can be used to achieve this based on two datasets: (i) *cirrhosis*, based on a study analyzing the differences in microbial compositions between $n = 130$ healthy and cirrhotic patients [Qin et al., 2014] and (ii) *centralparksoil*, based on a study analyzing the pH concentration using microbial compositions from $n = 580$ soil samples [Ramirez et al., 2014]. Our aim is not do draw novel biological conclusions, but rather to showcase how `KernelBiome` can be used in this type of analysis.

To reduce the complexity, we screen the data using `KernelBiome` with the Aitchison kernel and only keep the 50 taxa with the highest absolute CFIs. (As the analysis described here is only an illustration and we are not trying to compare methods, overfitting in this screening step is not a concern. In practice, however, it may be relevant to validate the sensitivity of the results using for example subsampling). We then fit `KernelBiome` with default parameter grid. For cirrhosis this results in the Aitchison kernel and for centralparksoil in the Aitchison-RBF kernel. As outlined in Section 2.2.3.2, we can then apply a kernel PCA with a compositionally adjusted component influence. The result for centralparksoil is given in Fig 2.3.5(a) (for cirrhosis see Fig 2.C.4 in Appendix 2.C). This provides some direct information on which perturbations affects each principle component (e.g., "[g]DA101[s]98" affects the first component the most positively and "Sphingobacterium[s]multivorum" affects the second component the most negatively). Moreover, it also directly provides a tool to detect groupings or outliers of the samples. For example, the samples in the top middle (i.e., center of the U-shape) in Fig 2.3.5(a) could be investigated further as they behave different to the rest.

A further useful quantity is the CFI, which for cirrhosis is given in Fig 2.3.5(b)

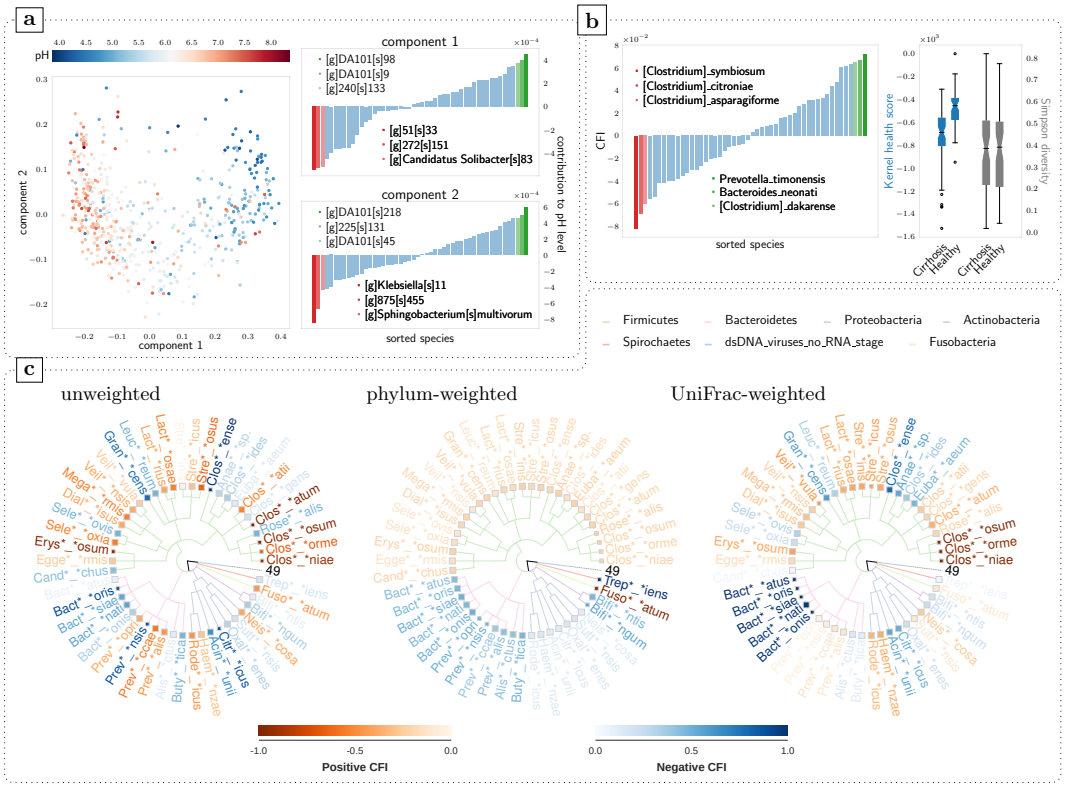



Figure 2.3.5: (a) shows a kernel PCA for the centralparksoil dataset with 2 principle components. On the right, the contribution of the species to each of the two components is given (see Appendix 2.E.2 for details). (b) and (c) are both based on the cirrhosis dataset. In (b) the CFI values are shown on the left and the right plot compares the proposed kernel health score with Simpson diversity. In (c) the scaled CFI values for are illustrated for different weightings. A darker color shade of the (shortened) name of the microbiota signifies a stronger (positive resp. negative) CFI.

(left) (for centralparksoil see Fig E in S3 Appendix). They explicitly take the compositional structure into account and have an easy interpretation. For example, "Prevotella_timonensis" has a CFI of 0.07 which implies that on average solely increasing "Prevotella_timonensis" will lead to a larger predicted response. We therefore believe that CFIs are more trustworthy than relying on for example Gini-importance for random forests, which does not have a clear interpretation due to the compositional constraint.

Lastly, one can also use the connection between kernels and distances to construct useful scalar summary statistics. In Fig 2.3.5(b) (right), we use the kernel-distance to the geometric median in the healthy subpopulation as a scalar indicator for the healthiness of the microbiome. In comparison with more standard scalar summary statistics such as the Simpson diversity, it is targeted to distinguish the two groups.

### 2.3.4 Model analysis given prior information

Including prior information in `KernelBiome` can be used to improve the interpretation of the model analysis step. To illustrate how this works in practice, we again consider the screened cirrhosis dataset. We apply KernelBiome with an Aitchison-kernel and $c$ equal to half the minimum non-zero relative abundance without weighting, with a phylum-weighting and with a UniFrac-weighting. The resulting scaled CFI values for each are visualized in Fig 2.3.5(c). The phylum-weighting corresponds to giving all taxa within the same phylum the same weights and the UniFrac-weighting is a weighting that incorporates the phylogenetic structure based on the UniFrac-distance and is described in Appendix 2.B.3. As can bee seen in Fig 2.3.5(c), the phylum weighting assigns approximately the same CFI to each variable in the same phylum, this is expected given that the phylum weighting has exactly the structure given in Proposition 2.2.4. Moreover, the UniFrac-weighting leads to CFI values that lie in-between the unweighted and phylum-weighted versions. Similar effects are seen for different kernels as well. The same plots for the generalized-JS kernel are provided in Fig 2.C.7 in Appendix 2.C.

## 2.4 Discussion and conclusions

In this work, we propose the `KernelBiome` framework for supervised learning with compositional covariates consisting of two main ingredients: data-driven model selection and model interpretation. Our approach is based on a flexible family of kernels targeting the structure of microbiome data, and is able to work with different kernel-based algorithms such as SVM and kernel ridge regression. One can also incorporate prior knowledge, which is crucial in microbiome data analysis. We compare `KernelBiome` with other state-of-the-art approaches on 33 microbiome datasets and show that `KernelBiome` achieves improved or comparable results. Moreover, `KernelBiome` provides multiple ways to extract interpretable information from the fitted model. Two novel measures, CFI and CPD, can be used to analyze how each component affects the response. We prove the consistency of these two measures and illustrate them on simulated and real datasets. `KernelBiome` also leverages the connection between kernels and distances to conduct distance-based analysis in a lower-dimensional space.

## Acknowledgements

The authors would like to thank Christian Müller for detailed feedback and suggestions on this work, Johannes Ostner for help creating the circle plots, Jeroen Raes and Doris Vandeputte for making their raw data available.

## Supporting information for "Supervised learning and model analysis with compositional data"

**Appendix 2.A Details on CFI and CPD.** Formal definitions of perturbations and estimators related to CFI and CPD.

**Appendix 2.B Details on kernels included in `KernelBiome`.** Overview of different kernel types, details on how they connect to distances and description of weighted kernels.

**Appendix 2.C Details and additional results for experiments in Section 2.3.** Datasets pre-processing, parameter setup, construction of the weighting matrices with UniFrac-distance and further experiment results based on the cirrhosis and centralpark-soil datasets.

**Appendix 2.D Additional experiments with simulated data.** Consistency of CFI and CPD and comparison of CFI and CPD with their non-simplex counterparts.

**Appendix 2.E Background on kernels.** Mathematical background on kernels and details on dimensionality and visualization with kernels.

**Appendix 2.F Proofs.** Proof of theorems and propositions.

**Appendix 2.G List of kernels implemented in `KernelBiome`.**

## 2.A. Details on CFI and CPD

### 2.A.1 Perturbations

Formally, the multiplicative perturbation $\psi$ and the fixed coordinate perturbation $\varphi$ are defined as follows.

- For all $j \in \{1, \dots, p\}$, $x \in \mathbb{S}^{p-1}$ with $x^j \neq 1$ and $c \in [0, \infty)$, define

$$\psi_j(x, c) := s_c(x^1, \cdots, x^{j-1}, cx^j, x^{j+1}, \cdots, x^p) \in \mathbb{S}^{p-1},$$

  where $s_c = 1/(\sum_{\ell \neq j}^p x^\ell + cx^j)$.

- For all $j \in \{1, \dots, p\}$, $x \in \mathbb{S}^{p-1}$ with $\sum_{\ell \neq j}^p x^\ell > 0$ and $c \in [0, 1]$, define the intervened composition by

$$\varphi_j(x, c) := (sx^1, \cdots, sx^{j-1}, c, sx^{j+1}, \cdots, sx^p) \in \mathbb{S}^{p-1},$$

  where $s = (1 - c)/(\sum_{\ell \neq j}^p x^\ell)$.

### 2.A.2 Estimators

We propose to estimate CFI and CPD with the following two estimators.

- For i.i.d. observations $X_1, \ldots, X_n$ and a differentiable function $f : \mathbb{S}^{p-1} \to \mathbb{R}$, we estimate the CFI for all $j \in \{1, \ldots, p\}$ as

$$\hat{I}_f^j = \frac{1}{n} \sum_{i=1}^{n} \tfrac{d}{dc} f(\psi(X_i, c))\big|_{c=1}.$$

- For i.i.d. observations $X_1, \ldots, X_n$ and a function function $f : \mathbb{S}^{p-1} \to \mathbb{R}$, we estimate the CPD for all $j \in \{1, \ldots, p\}$ and $z \in [0, 1]$ as

$$\hat{S}_f^j(z) = \frac{1}{n} \sum_{i=1}^{n} f(\varphi(X_i, z)) - \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

## 2.B. Details on kernels included in `KernelBiome`

### 2.B.1 Overview of kernels in `KernelBiome`

In this section, we give additional details on the kernels used in `KernelBiome`. A full list of all kernels and their corresponding metrics together with a visualization on $\mathbb{S}^2$ is given in Appendix 2.G.

As discussed in the main paper, we consider four types of kernels.

- **Euclidean** These are kernels that are used on Euclidean space but restricted to the simplex. This includes the *linear kernel* and the *RBF kernel*.
- **Probability distribution** These are kernels that are constructed from metrics between probability distributions. `KernelBiome` includes two parametric classes of kernels, the *Hilbertian kernel* and the *generalized-JS kernel*. These kernels correspond to multiple well-known metrics on probabilities such as the *chi-squared metric*, the *total-variation metric*, the *Hellinger metric* and the *Jensen-Shannon metric*.
- **Aitchison geometry** These are kernels that are constructed by using the centered log-ratio transform to project data on the simplex into Euclidean space and then combining it with a Euclidean kernel. `KernelBiome` includes the *Aitchison kernel* and the *Aitchison RBF kernel*. In order to allow for zeros, a small positive number $c$ is added to each coordinate for all observations before applying the centered log-ratio transformation.
- **Riemannian manifold** These kernels are connected to the simplex via multinomial distributions and have been shown to empirically perform well on sparse text data mapped into the simplex. `KernelBiome` contains the *heat-diffusion kernels*.

For each type of kernel there are multiple parameter settings. Although users of the `KernelBiome` package can freely change the parameters, the default settings for `KernelBiome` for each type of kernel are provided by the package and are given in Table 2.B.1.

| Geometry | Kernel | Parameters | Number of kernels |
|---|---|---|---|
| Euclidean | linear | none | 1 |
| | RBF | $\sigma^2 \in \{10^{-2} \cdot m_1, 10^{-1} \cdot m_1, m_1, 10 \cdot m_1, 10^2 \cdot m_1, 10^3 \cdot m_1, 10^4 \cdot m_1\}$ | 7 |
| Probability distribution | generalized-JS | $(a, b) \in \{(1, 0.5), (1, 1), (10, 0.5), (10, 1), (10, 10), (\infty, 0.5), (\infty, 1), (\infty, 10), (\infty, \infty)\}$ | 9 |
| | Hilbertian | $(a, b) \in \{(1, -1), (1, -10), (1, -\infty), (10, -1), (10, -10), (10, -\infty), (\infty, -1), (\infty, -10)\}$ | 8 |
| Aitchison geometry | Aitchison | $c \in \{\mu_X/2 \cdot 10^{-4}, \ldots, \min(\mu_X/2 \cdot 10^4, 10^{-2})\}$ | 9 |
| | Aitchison-RBF | $c \in \{\mu_X/2 \cdot 10^{-4}, \ldots, \min(\mu_X/2 \cdot 10^4, 10^{-2})\}$, $\sigma \in \{c \cdot m_2 \cdot 10^{-1}, c \cdot m_2, c \cdot m_2 \cdot 10\}$ | 15 |
| Riemannian manifold | heat-diffusion | $t = x^{\frac{2}{n-1}} \frac{1}{4\pi}$ for $x \in \{10^{-20}, \ldots, 10\}$ | 6 |

Table 2.B.1: Default parameter grid in `KernelBiome`. $m_1$ and $m_2$ are the median heuristic for the RBF and Aitchison-RBF kernel, respectively, which depend on the data. $\mu_X$ is the minimal non-zero value in $X$. The zero grids for the Aitchison geometry kernels have an even logarithmic spacing and contain 9 and 5 parameters for the Aitchison and Aitchison-RBF, respectively. Similarly, the grid for $x$ for the heat-diffusion kernel has an even logarithmic spacing with 6 values. There are a total of 55 kernels.

## 2.B.2 Connecting positive definite kernels to metrics

A semi-metric $d$ satisfies all properties of a metric, except that $d(x, y) = 0$ does not imply $x = y$. This can happen because a kernel can map two different points in $\mathcal{X}$ to the same point in $\mathcal{H}_k$. Any fixed kernel $k$ on $\mathcal{X}$ induces a semi-metric $d_k$ on $\mathcal{X}$ defined for all $x, y \in \mathcal{X}$ by

$$d_k^2(x, y) = k(x, x) + k(y, y) - 2k(x, y). \tag{2.B.1}$$

This holds for all positive-definite kernels by Theorem 2.E.7 in Appendix 2.E. In particular, this corresponds to the distance between the embedded points in the RKHS $\mathcal{H}_k$, that is,

$$\|k(x, \cdot) - k(y, \cdot)\|_{\mathcal{H}_k} = d_k(x, y).$$

The feature embedding $x \mapsto k(x, \cdot)$ induced by a kernel therefore preserves the distances $d_k$. A useful aspect of kernel methods, is that they allow a post-analysis based on the embedded features, see also Section 2 in S5 Appendix.

A partial reverse implication is also true. For a particular type of semi-metric $d$ on $\mathcal{X}$ (these metrics are called Hilbertian, see Appendix 2.E) it is possible to construct a

kernel $k$ on $\mathcal{X}$ defined for all $x, y \in \mathcal{X}$ by

$$k(x,y) = -\tfrac{1}{2}d^2(x,y) + \tfrac{1}{2}d^2(x,x_0) + \tfrac{1}{2}d^2(x_0,y),$$

where $x_0 \in \mathcal{X}$ is an arbitrary reference point, such that the distance in the corresponding RKHS $\mathcal{H}_k$ is $d$.

Kernels can be shifted in such a way that the origin in the induced RKHS changes but the metric in (2.B.1) remains fixed (see Lemma 2.E.6 in S5 Appendix). A natural origin in the simplex is given by the point $u = (\frac{1}{p}, \ldots, \frac{1}{p})$, therefore we have shifted all kernels such that $k(u, \cdot) \equiv 0$ and hence correspond to the origin in $\mathcal{H}_k$. In S5 Appendix, we provide a short overview of the mathematical results that connect kernels and metrics.

### 2.B.3 Weighted kernels - including prior information

In this section, we discuss how to include prior knowledge, e.g. phylogenetic information, into the simplex kernels. We assume the information is encoded in a matrix $W \in \mathbb{R}^{p\times p}$ where each element corresponds to a measure of similarity between components. That is, $W_{i,j}$ is large if components $i$ and $j$ are similar (or related) and small otherwise. We assume that $W$ is symmetric, positive semi-definite and all entries in $W$ are non-negative.

The linear kernel and all kernels based on probability distributions have the form

$$k(x,y) = \textstyle\sum_{i=1}^{p} k_0(x^i, y^i) \tag{2.B.2}$$

and we therefore define the weighted version by

$$k_W(x,y) = \sum_{j,\ell=1}^{p} W_{j,\ell} \cdot k_0(x^j, y^\ell). \tag{2.B.3}$$

The weighted versions of the remaining kernels are defined individually. A full list of the weighted kernels is given in Appendix 2.G.2.

#### 2.B.3.1 Validity of weighted kernels

In order to use the proposed weighted kernels, we need to ensure that they are indeed positive definite. In the following, we prove this for the weighted versions of the *linear kernel*, the *Hilbertian kernel*, the *Generalized-JS kernel*, the *RBF kernel* and the *Aitchison kernel*. We do not prove it for the *Aitchison RBF kernel* and the *Heat Diffusion kernel* and only note that they appear to be positive definite from our empirical evaluations.

We begin by showing that the kernel defined in (2.B.3) is positive definite whenever $k_0 : [0,1] \times [0,1] \to \mathbb{R}$ is positive definite. To see this, fix $x_1, \ldots, x_n \in \mathbb{S}^{p-1}$ and $\alpha \in \mathbb{R}^n$ and denote by $K_W \in \mathbb{R}^{n\times n}$ the kernel Gram-matrix based on $x_1, \ldots, x_n$ and kernel $k_W$.

Then,

$$\begin{aligned}\alpha^\top K_W \alpha &= \sum_{i,r=1}^{n}\sum_{j,\ell=1}^{p} \alpha_i \alpha_r W_{j,\ell} k_0(x_i^j, x_r^\ell) \\ &= \sum_{j,\ell=1}^{p} W_{j,\ell} \left( \sum_{i,r=1}^{n} \alpha_i \alpha_r k_0(x_i^j, x_r^\ell) \right).\end{aligned}$$

Since $k_0$ is positive definite, it holds that $\sum_{i,r} \alpha_i \alpha_r k_0(x_i^j, x_r^\ell) \geq 0$ and hence $\alpha^\top K_W \alpha \geq 0$ since all entries in $W$ are non-negative.

We now go over the individual weighted kernels and argue that they are positive definite.

- **Linear kernel** Since $\mathbb{R}$ is a Hilbert space with the inner product $xy$ which induces the $|x-y|$ it follows that the squared distance $d^2_{\text{Linear}}(x,y) \coloneqq (x-y)^2$ is Hilbertian as well. Applying Theorem 2.E.3 in Appendix 2.E we know that the distance is of negative type. Thus, based on the one-dimensional squared linear distance $d^2_{\text{Linear}}$, we apply Theorem 1.2 in Appendix 2.E with $x_0 = \frac{1}{p}$ to construct the following positive definite kernel $k_0$ defined for all $x, y \in [0,1]$ by

  $$k_0(x,y) \coloneqq -\tfrac{1}{2}(x-y)^2 + \tfrac{1}{2}(x-\tfrac{1}{p})^2 + \tfrac{1}{2}(\tfrac{1}{p}-y)^2 = xy - \tfrac{x}{p} - \tfrac{y}{p} + \tfrac{1}{p^2}.$$

  Comparing this with our weighted linear kernel in Appendix 2.G.2, we see that the weighted linear kernel has the form (2.B.3) and is therefore positive definite by the above argument.

- **Hilbertian kernel** As shown by Hein and Bousquet [2005] the distance $d_{\text{Hilbert}} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined for all $x, y \in \mathbb{R}_+$ by

  $$d^2_{\text{Hilbert}}(x,y) = \frac{2^{\frac{1}{b}}\left[x^a + y^a\right]^{\frac{1}{a}} - 2^{\frac{1}{a}}\left[x^b + y^b\right]^{\frac{1}{b}}}{2^{\frac{1}{a}} - 2^{\frac{1}{b}}}$$

  is a Hilbertian metric on $\mathbb{R}_+$. Applying Theorem 2.E.7 in Appendix 2.E with $x_0 = \frac{1}{p}$ results in a positive definite kernel $k_0$ that when combined as in (2.B.3) results in the proposed weighted Hilbertian kernels in Appendix 2.G.2. Therefore, we have shown that the weighted Hilbertian kernels are positive definite as long as $W$ has non-negative entries.

- **Generalized-JS kernel** Similarly the weighted Generalized-JS kernels in Appendix 2.G.2x can all be decomposed as in (2.B.3) with a one-dimensional kernels $k_0$ on $[0,1]$. Topsøe [2003] show that all these $k_0$ can be generated using Theorem 2.E.7 in Appendix 2.E with $x_0 = \frac{1}{p}$ based on Hilbertian metrics. Hence, all weighted Generalized-JS kernels are positive definite as long as $W$ has non-negative entries.

- **Aitchison kernel** To show that the weighted Aitchison kernel (defined in Appendix 2.G.2) is positive definite, we first define the mapping $\Phi : \mathbb{S}^{p-1} \to \mathbb{R}^p$ by $\Phi(x) := \frac{x+c}{g(x+c)}$. Then, the weighted Aitchison kernel is given by

$$k(x, y) = \Phi(x)^\top W \Phi(y).$$

Since $W$ is symmetric and positive semi-definite there exists $M \in \mathbb{R}^{p\times p}$ such that $W = M^\top M$. Therefore, for any $\alpha \in \mathbb{R}^n$ and $x_1, \dots, x_n \in \mathbb{S}^{p-1}$ it holds that

$$\sum_{i,r} \alpha_i \alpha_r k(x_i, x_r) = \sum_{i,r} \alpha_i \alpha_r (M\Phi(x_i))^\top M\Phi(x_r) \geq 0.$$

Hence, $k$ is positive definite.

- **RBF kernel** Using the symmetry of $W$ the weighted RBF kernel can be expressed as follows

$$\begin{aligned} k(x, y) &= \exp\Big( -\frac{1}{\sigma^2} \sum_{j,\ell=1}^{p} W_{j,\ell}(x^j - y^\ell)^2 \Big) \\ &= \underbrace{\exp\Big( -\frac{1}{\sigma^2} \sum_{j,\ell=1}^{p} W_{j,\ell}(x^j)^2 \Big) \exp\Big( -\frac{1}{\sigma^2} \sum_{j,\ell=1}^{p} W_{j,\ell}(y^j)^2 \Big)}_{=:A(x,y)} \\ &\quad \cdot \underbrace{\exp\Big( \frac{2}{\sigma^2} \sum_{j,\ell=1}^{p} W_{j,\ell} x^j y^\ell \Big)}_{=:B(x,y)} \end{aligned}$$

The function $A$ is a positive definite kernel since it is the inner-product of a feature mapping. The function $B$ can be shown to be a kernel by considering the Taylor expansion of the exponential function and using that sums and limits of positive definite kernels are again positive definite together with the fact that $W$ is positive semi-definite. Therefore, the weighted RBF kernel is positive definite.

### 2.B.4 UniFrac-Weighting

In this section, we show how prior information based on the UniFrac-Distance [Lozupone and Knight, 2005] can be encoded into a weight matrix $W \in \mathbb{R}^{p\times p}$. Depending on the application at hand different distances can be used in a similar way. The UniFrac-Distance is a $\beta$-diversity measure that uses phylogenetic information to compare two compositional samples $x, y \in \mathbb{S}^{p-1}$. Each element of the sample is hereby placed on a phylogenetic tree. The distance between both samples is computed via quantification of overlapping branch length, that is,

$$\text{UniFrac-Distance}(x, y) = \frac{\text{sum of unshared branch length of } x \text{ and } y}{\text{sum of all tree branch length of } x \text{ and } y} \in [0, 1].$$

Based on the UniFrac-Distance, we define two similarity matrices $M^A, M^B \in [0,1]^{p\times p}$ for all $i,j \in \{1,\dots,p\}$ by

$$M^A_{i,j} \coloneqq 1 - \text{UniFrac-Distance}(e_i, e_j),$$
$$M^B_{i,j} \coloneqq \sum_{\ell=1}^{p} \text{UniFrac-Distance}(e_i, e_\ell) \cdot \text{UniFrac-Distance}(e_j, e_\ell),$$

where $e_i, e_j \in \mathbb{S}^{p-1}$ with 1 on the $i$-th and $j$-th coordinate, respectively. $M^A$ and $M^B$ are two options of encoding the UniFrac-Distance as a similarity. $M^B$ is positive semi-definite by construction, while this is not true for $M^A$ and should be checked empirically. We recommend using $M^A$ whenever it is positive semi-definite.

We then construct the weight matrix $W^{\text{UniFrac}} \in \mathbb{R}^{p\times p}$ by scaling $M^*$ such that the diagonal entries are one, that is,

$$W^{\text{UniFrac}} \coloneqq DM^*D,$$

where $D = \text{diag}(\sigma_1,\dots,\sigma_p)$, with $\sigma_i = 1/\sqrt{M^*_{i,i}}$. Since by construction the matrix $M^*$ has its largest values on the diagonal, this weight matrix takes values in $[0,1]$. Moreoever, by construction it remains symmetric and positive semi-definite.

## 2.C. Details and additional results for experiments

### 2.C.1 List of datasets

| **Dataset** | **Reference** | **Prediction tasks** | **Dim** ($n \times d$) | **Additional preprocessing** |
|---|---|---|---|---|
| rmp | Vandeputte et al. [2017] | classification:<br>- Crohn's disease vs healthy | $95 \times 351$ | none |
| camp | Berry et al. [2020] | classification:<br>- parasite infected vs healthy | $270 \times 622$ | none |
| cirrhosis | Qin et al. [2014] | classification:<br>- cirrhosis vs healthy | $130 \times 444$ | aggregated to species & prev./abun. filtering |
| cancer | Baxter et al. [2016] | classification:<br>- cancer vs non-cancer | $490 \times 335$ | none |
| impaired-diabetes | Karlsson et al. [2013] | classification:<br>- impaired vs type 2 diabetes | $101 \times 3758$ | none |
| nugent-category | Ravel et al. [2011] | classification:<br>- nugent score high vs low | $342 \times 305$ | none |
| gastro-oral | Human Microbiome Project Consortium [2012] | classification:<br>- gastrointestinal vs oral | $2070 \times 1218$ | none |

| | | | | |
|---|---|---|---|---|
| healthy-cd | Morgan et al. [2012] | classification:<br>- healthy vs Crohn's disease | $74 \times 367$ | none |
| kostic | Kostic et al. [2012] | classification:<br>- healthy vs tumor | $172 \times 409$ | none |
| malawi-venezuela | Yatsunenko et al. [2012] | classification:<br>- Malawi vs Venezuela | $54 \times 1544$ | none |
| black-hispanic | Ravel et al. [2011] | classification:<br>- black vs Hispanic | $199 \times 305$ | none |
| ss-paired | Human Microbiome Project Consortium [2012] | classification:<br>- sub vs supragingival plaque | $408 \times 1218$ | none |
| usa-malawi | Yatsunenko et al. [2012] | classification:<br>- US vs Malawi | $150 \times 1544$ | none |
| st-paired | Human Microbiome Project Consortium [2012] | classification:<br>- stool vs tongue dorsum | $404 \times 1218$ | none |
| gevers_ileum | Gevers et al. | classification:<br>- Crohn's disease vs healthy | $140 \times 446$ | none |
| yatsunenko_sex | Yatsunenko et al. [2012] | classification:<br>male vs female | $129 \times 1544$ | none |
| normal-diabetes | Karlsson et al. [2013] | classification:<br>- normal vs type 2 diabetes | $96 \times 3758$ | none |
| healthy-uc | Morgan et al. [2012] | classification:<br>- healthy vs Ulcerative colitis | $59 \times 367$ | none |
| hmp_sex | Human Microbiome Project Consortium [2012] | classification:<br>- female vs male | $180 \times 1218$ | none |
| qin2012 | Qin et al. [2012] | classification:<br>- healthy vs type 2 diabetes | $124 \times 2526$ | none |
| turnbaugh | Turnbaugh et al. [2007] | classification:<br>- lean vs obese | $142 \times 232$ | none |
| gevers_rectum | Gevers et al. | classification:<br>- Crohn's disease vs health | $160 \times 446$ | none |
| qin2014 | Qin et al. [2014] | classification:<br>- cirrhosis vs healthy | $130 \times 2579$ | none |
| white-black | Ravel et al. [2011] | classification:<br>- white vs black | $200 \times 305$ | none |
| centralparksoil | Ramirez et al. [2014] | regression:<br>- ph level of soil | $580 \times 1498$ | prev./abun. filtering |
| uk | McDonald et al. [2018] | regression:<br>- BMI | $882 \times 327$ | UK subpopulation & prev./abun. filtering |
| hiv | Rivera-Pinto et al. [2018] | regression:<br>- CD4+ cell counts | $152 \times 282$ | none |
| tara | Sunagawa et al. [2020] | regression:<br>- ocean salinity | $136 \times 2407$ | prev./abun. filtering |
| ravel_ph | Ravel et al. [2011] | regression:<br>- vaginal pH | $388 \times 305$ | none |

| | | | | |
|---|---|---|---|---|
| pcdai-rectum | Gevers et al. | regression: - PCDAI scores | $51 \times 446$ | none |
| pcdai-ileum | Gevers et al. | regression: - PCDAI scores | $67 \times 446$ | none |
| baby-age | Yatsunenko et al. [2012] | regression: - infant age | $49 \times 1544$ | none |
| nugent-score | Ravel et al. [2011] | regression: - nugent score | $388 \times 305$ | none |

Table 2.C.1: List of microbiome datasets used to benchmark `KernelBiome`. The cirrhosis dataset is a processed version of qin2014. Whenever prevalence/abundance filtering (prev./abun. filtering) is applied it means that only taxa that appear in 25% of the samples and with a median non-zero count of 5. Datasets other than the following ones are taken from the MLRepo [Vangay et al., 2019] are taken directly without further processing: uk, camp, centralparksoil, cirrosis, cancer, hiv, rmp, tara.

### 2.C.2 Weighting matrix for weighted `KernelBiome`

The weight matrix $W^{\text{UniFrac}}$ for the cirrhosis dataset [Qin et al., 2014] and the centralparksoil dataset [Ramirez et al., 2014] are presented as heatmaps in Fig 2.C.1. Using our proposed weighted kernels (see Appendix 2.G.2) with the UniFrac-based weight matrix $W^{\text{UniFrac}}$ is different from incorporating the UniFrac-distance via kernel convolution as proposed by Zhao et al. [2015b].

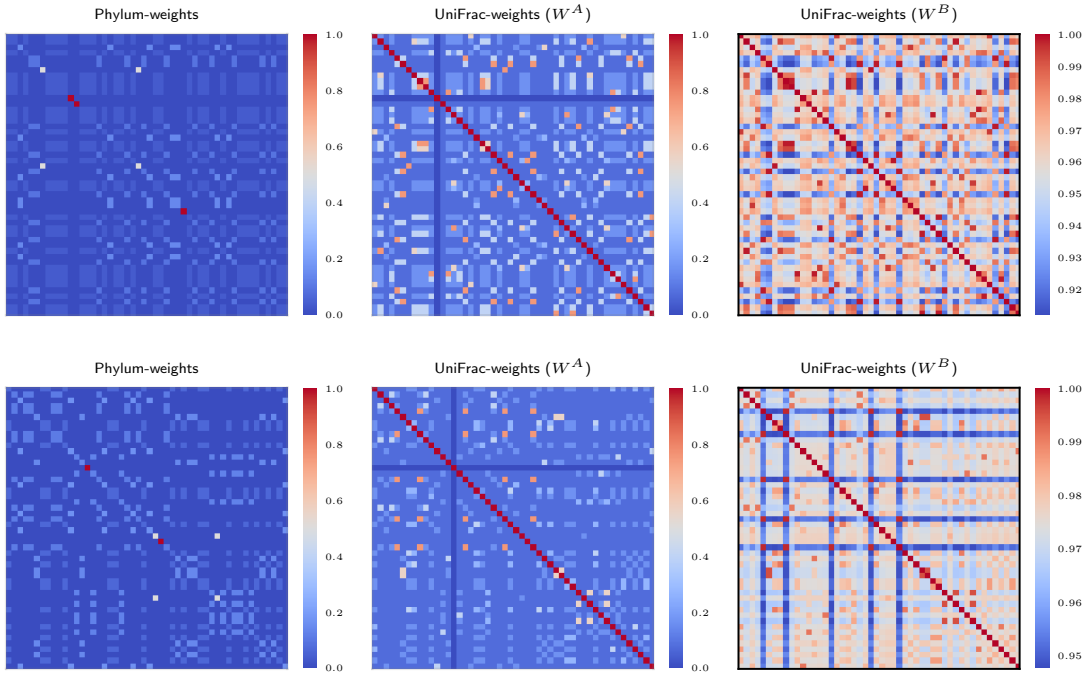



Figure 2.C.1: Visualization of the phylum-weights and the two UniFrac-weights $W^A = DM^AD$ for $W^B = DM^BD$, based on the 50 pre-screened species (see Appendix 2.C). Upper panel: cirrhosis dataset. Lower panel: centralparksoil dataset.

## 2.C.3 Detailed experiment results on public datasets

### 2.C.3.1 Prediction performance

Here we provide more details on the the prediction performance evaluation. Boxplots of prediction scores for all 33 datasets as in Fig 2.3.3 in the main text are given in Fig 2.C.2, and PR curves accompanying the boxplots can be found in Fig 2.C.3. We can see that `KernelBiome` achieves the best results for most of the tasks. For classification tasks, `KernelBiome` performs competitively both in terms of balanced accuracy and PR curves. (For `SVM-RBF`, `KB-Aitchison` and `KernelBiome`, the PR curves are based on the estimated probabilities computed in the sklearn-package. We observed a slight mismatch between these predicted probabilities and predicted classes in some of the examples, which is due to a bug `https://github.com/scikit-learn/scikit-learn/issues/13211`. We therefore recommend putting more emphasis on the accuracy plots.) The frequency of kernels selected the most often by `KernelBiome` is given in Table. 2.C.2.

| Dataset (short name) | Kernel | Frequency (%) |
|---|---|---|
| rmp | aitchison | 59.5 |
| camp | aitchison-rbf | 44.5 |
| cirrhosis | aitchison-rbf | 65.5 |
| cancer | aitchison | 73.5 |
| impaired-diabetes | aitchison | 54.0 |
| nugent-category | aitchison-rbf | 63.0 |
| gastro-oral | aitchison | 100.0 |
| healthy-cd | aitchison-rbf | 46.0 |
| kostic | aitchison-rbf | 55.5 |
| malawi-venezuela | aitchison | 100.0 |
| black-hispanic | aitchison | 55.5 |
| ss-paired | aitchison-rbf | 78.5 |
| usa-malawi | aitchison | 100.0 |
| st-paired | aitchison | 100.0 |
| gevers_ileum | generalized-js | 29.5 |
| yatsunenko_sex | aitchison-rbf | 70.5 |
| normal-diabetes | aitchison-rbf | 48.0 |
| healthy-uc | aitchison-rbf | 48.5 |
| hmp_sex | aitchison-rbf | 35.5 |
| qin2012 | aitchison | 53.0 |
| turnbaugh | aitchison | 56.5 |
| gevers_rectum | aitchison-rbf | 53.5 |
| qin2014 | aitchison-rbf | 81.5 |
| white-black | aitchison | 50.0 |
| centralparksoil | generalized-js | 45.0 |
| uk | aitchison-rbf | 100.0 |
| hiv | aitchison-rbf | 97.0 |
| tara | aitchison-rbf | 93.0 |
| ravel_ph | heat-diffusion | 61.0 |
| pcdai-rectum | rbf | 67.0 |
| pcdai-ileum | rbf | 40.0 |
| baby-age | aitchison | 91.0 |
| nugent-score | aitchison-rbf | 99.5 |

Table 2.C.2: Kernels selected most frequently by `KernelBiome` for all 33 datasets.

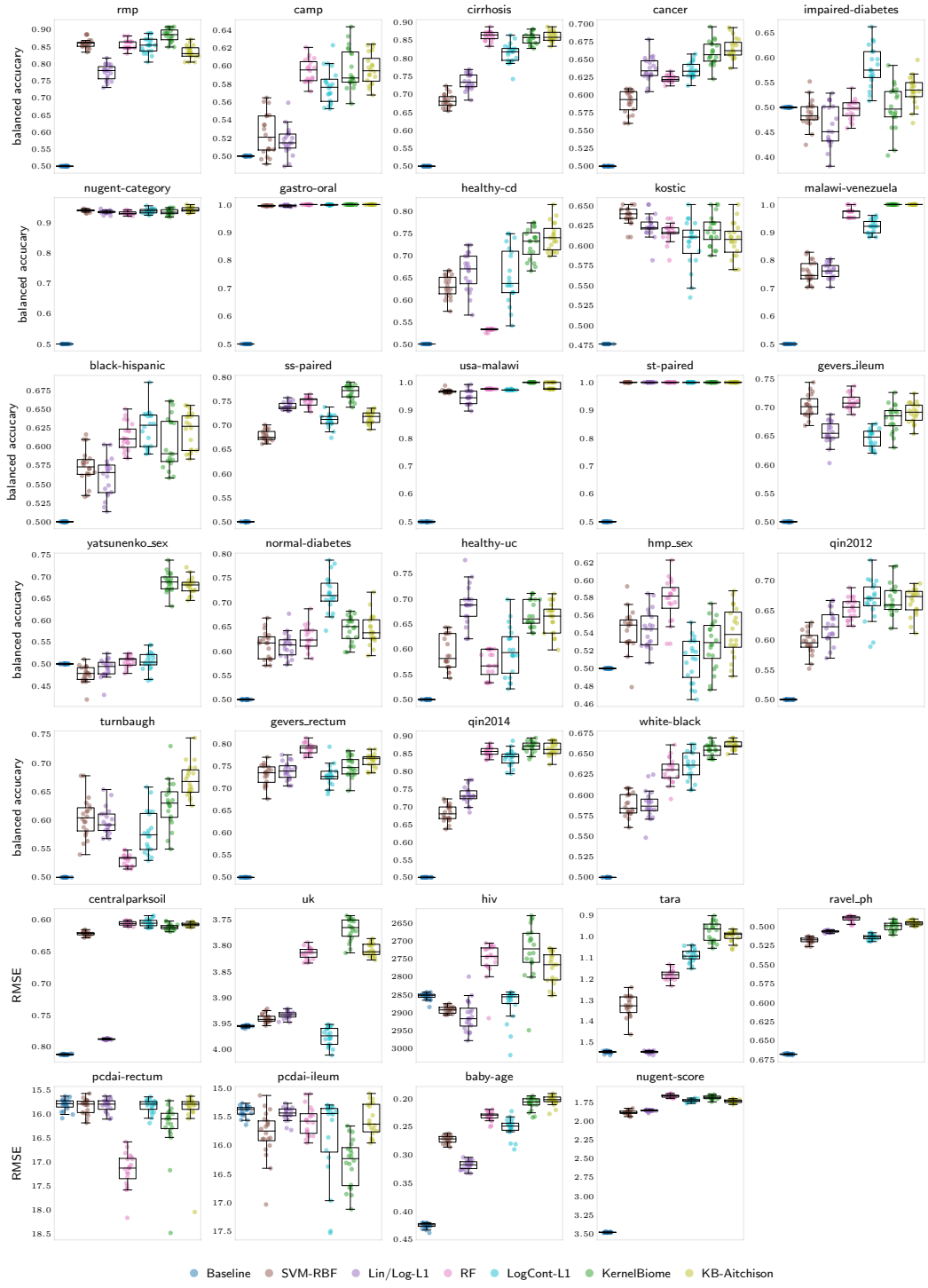



Figure 2.C.2: Comparison of predictive performance on the 33 public datasets based on a 10-fold train/test split.

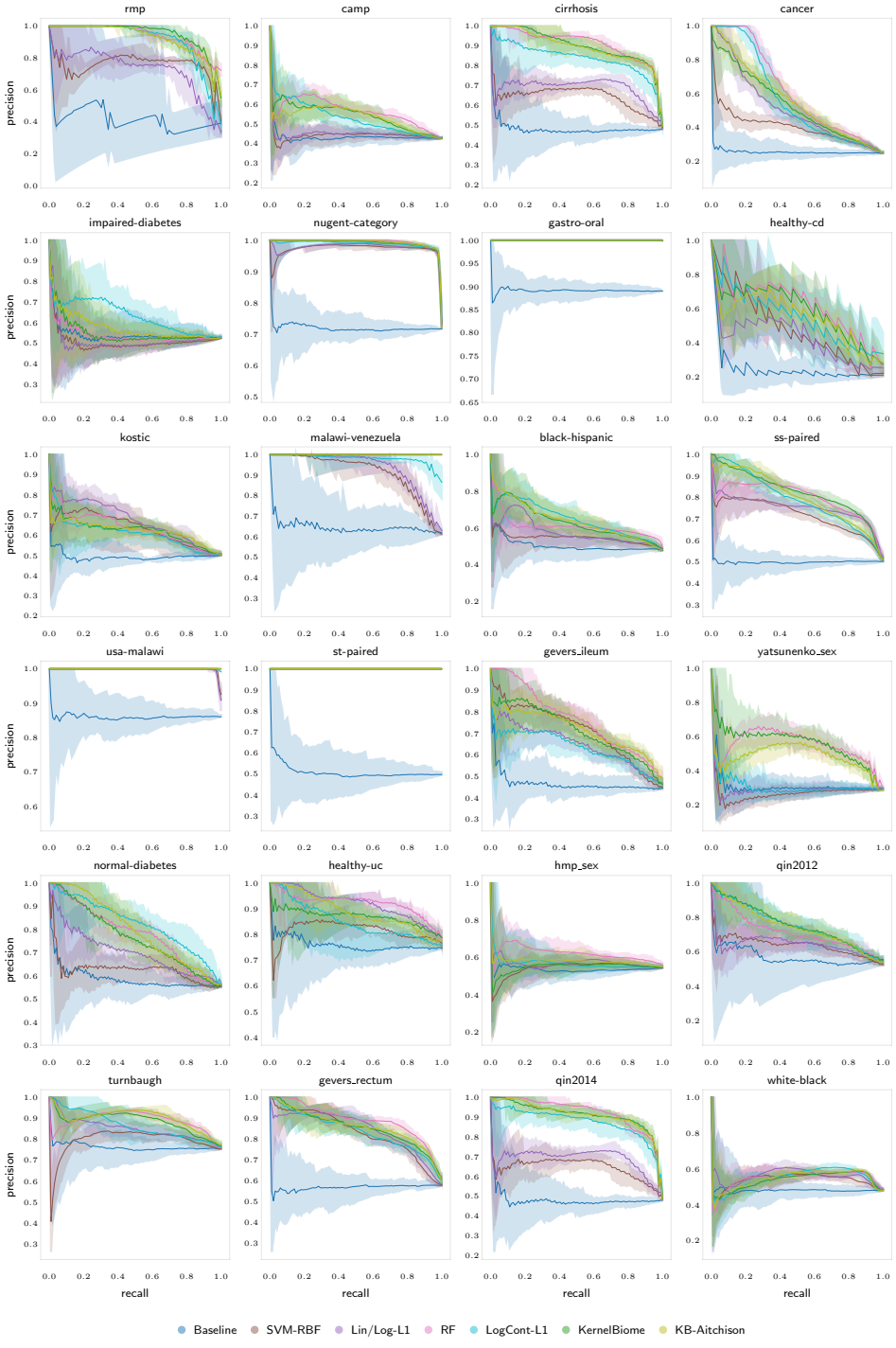



Figure 2.C.3: PR curves for the 24 classification datasets. The solid curve is the average curve from the 20 random 10-fold CV, and the shaded area is the 95% confidence band.

### 2.C.3.2 Model analysis

Here we include the remaining model analysis results for the cirrhosis and centralparksoil datasets. As in the main paper, we screened both data sets to only include the 50 taxa with the highest absolute CFIs by `KernelBiome` with Aitchison kernel. Kernel PCA plots for the cirrhosis dataset is given in Fig 2.C.4 and the CFI values for the centralparksoil dataset are given in Fig 2.C.5.

Furthermore, we also provide the missing circle plots here. The extended version of Fig 5C in the main text with long labels is given in Fig 2.C.6. Fig 2.C.7 is the circle plot for the cirrhosis dataset based on the generlized-JS kernel. Fig 2.C.8is the circle plot for the centralparksoil dataset based on the Aitchison kernel.

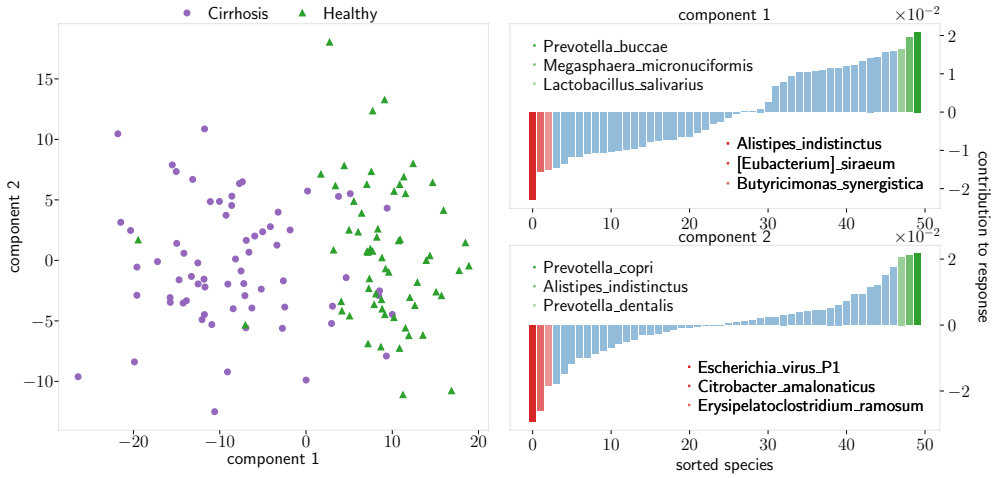



Figure 2.C.4: Kernel PCA plot and contributions of the 50 taxa to component 1 and 2 sorted from the most negative contribution to the most positive contribution (cirrohsis).

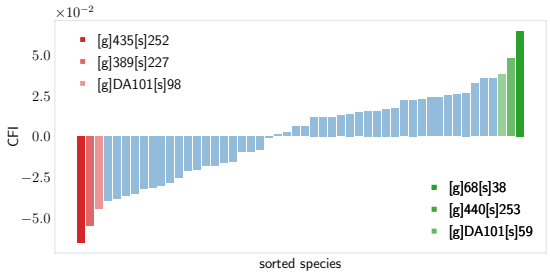



Figure 2.C.5: CFI values for the 50 taxa sorted from the most negative contribution to the most positive contribution to the response (centralparksoil).

## 2.D. Additional experiments with simulated data

### 2.D.1 Consistency of CPD and CFI

We illustrate the consistency of CPD and CFI from Theorem 2.1 in the main text based on `KernelBiome` with the following example. Let $k_{\text{tv}}$ be the total variation kernel and consider the function

$$f : x \mapsto 100 \cdot k_{\text{tv}}(z, x)$$

with

$$z = (0.06544714, 0.08760064, 0.17203408, 0.07502236, 0.1642615,$$
$$0.03761901, 0.18255478, 0.13099514, 0.08446536) \in \mathbb{S}^8$$

being a fixed and randomly selected point. Furthermore, we generate an i.i.d. dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ based on the following 2 step generative model.

- **Step 1:** Generate a random variable $\tilde{X} = (\tilde{X}^1, \ldots, \tilde{X}^9)$ such that the three blocks $(\tilde{X}^1, \tilde{X}^2, \tilde{X}^3)$, $(\tilde{X}^4, \tilde{X}^5, \tilde{X}^6)$, and $(\tilde{X}^7, \tilde{X}^8, \tilde{X}^9)$ are i.i.d. from $\text{LogNormal}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & 0.25 & -0.25 \\ 0.25 & 1 & 0.25 \\ -0.25 & 0.25 & 1 \end{pmatrix}$. Then, $X_i$ is constructed by normalizing $\tilde{X}$, that is, $X_i = \tilde{X} / \sum_{j=1}^{9} \tilde{X}^j$. The block structure adds non-trivial correlation structure between the compositional components.

- **Step 2:** Generate $Y_i$ based on $X_i$ by

  $$Y_i = f(X_i) + \epsilon_i$$

  with $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

Based on one such dataset, we estimate the CFI and CPD for a fitted `KernelBiome` estimator (using kernel ridge regression and default settings), and compare the estimates against the population CFI and CPD calculated from the true function $f$. In Fig 2.D.1, we report the mean squared deviations (MSD) for both CFI and CPD based on 100 such datasets for each sample size.

### 2.D.2 Comparing CFI and CPD with permutation importance and partial dependence plots

Two common approaches to assess the importance of individual features are permutation importance (PI) and partial dependency plot (PDP). PI of the $j$-th feature is defined as the mean difference between the baseline mean squared error of a fitted model and the average mean squared error after permuting the $j$-th feature column a certain number of times. PDP is used to describe how individual features contribute to a fitted model. For the $j$-th feature, it describes its contribution by the function

| | $f_1$ | | | $f_2$ | | |
|---|---|---|---|---|---|---|
| | $x^1$ | $x^2$ | $x^3$ | $x^1$ | $x^2$ | $x^3$ |
| CFI | 0.85 | 0.87 | **-1.72** | 1.94 | -1.94 | **0.00** |
| RI | 3.76 | 2.99 | **0.00** | 0.00 | -4.72 | **-4.40** |
| PI | 11.66 | 5.76 | **0.00** | 0.00 | 28.98 | **24.72** |

Table 2.D.1: CFI, RI and PI for the two functions $f_1$ and $f_2$ defined in (2.D.1). Only CFI correctly attributes the effect of $x^3$ (marked in bold).

$z \mapsto \mathbb{E}[\hat{f}(X^1, \dots, X^{j-1}, z, X^j, \dots, X^p)]$, where $\hat{f}$ is the fitted model. Both PI and PDP can be misleading when used with compositional covariates.

In this section, we illustrate this based on two examples. In both cases, the proposed adjusted measures CFI and CPD remain correct, while the PI and PDP are incorrect. Consider the two functions

$$\begin{aligned} f_1 &: x \mapsto 10x^1 + 10x^2 \\ f_2 &: x \mapsto \frac{1 - x^2 - x^3}{1 - x^3}. \end{aligned} \tag{2.D.1}$$

For $f_1$, changes in all coordinates affect the function value due to the simplex constraint. For $f_2$, only changes in $x^1$ and $x^2$ affect the function value but not changes in $x^3$. This is because on the simplex $f_2(x) = \frac{x^1}{x^1+x^2}$. An importance measure should therefore associate a non-zero value to $x^3$ for $f^1$ and zero to $x^3$ for $f^2$.

We generate 200 i.i.d. observations $X_1 \dots, X_{100}$ with $X_i \overset{d}{=} \tilde{X}_i / \sum_{j=1}^3 \tilde{X}_i^j$ for $\tilde{X}_i \overset{\text{i.i.d.}}{\sim} \text{LogNormal}(0, \text{Id}_3)$ ($\text{LogNormal}(\mu, \Sigma)$ denotes the log-normal distribution with location parameter $\mu$ and scale parameter $\Sigma$. $\text{Id}_3$ denotes the 3-dimensional identity matrix.) and compute PI, PDP, CFI and CPD for each of the two functions. The results are given in Table 2.D.1 and Fig 2.D.2.

As expected, the CFI and also CPD correctly capture the behavior of the two functions. However, PI and PDP are incorrect in both cases: For $f_1$ the variable $x^3$ shows no effect both with PI and PDP and for $f_2$ the variable $x^3$ is falsely assigned a strong negative effect even though it does not affect the function value at all. In Table 2.D.1, we have additionally computed the relative influence (RI) given by $\mathbb{E}[\frac{d}{dx^j}\hat{f}(X)]$ due to Friedman [2001]. It has the same problems as PI as it does not take into account the simplex structure.

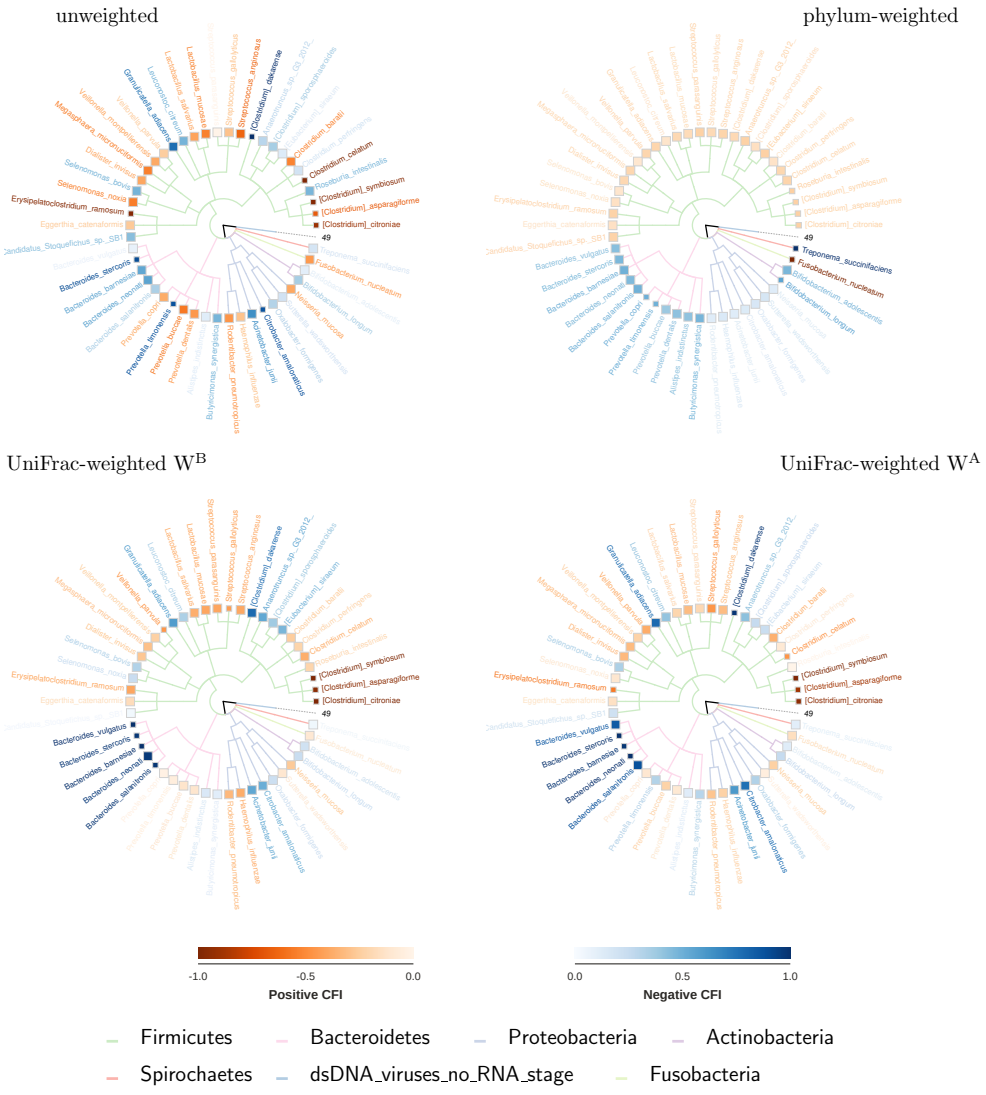



Figure 2.C.6: Scaled CFI values for cirrhosis dataset where a darker color shade of the name of the microbiota signifies a stronger (positive resp. negative) feature influence (Aitchison Kernel).

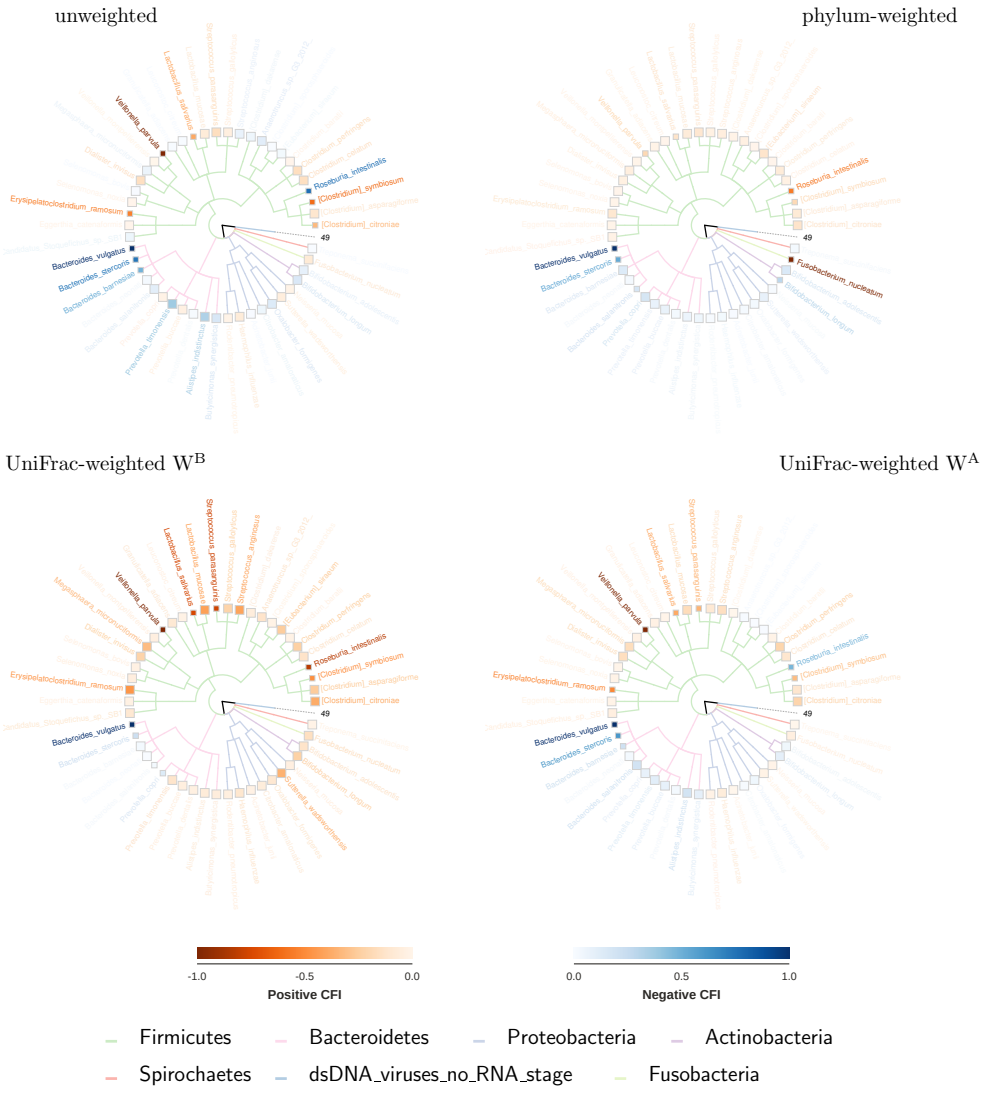



Figure 2.C.7: Scaled CFI values for cirrhosis dataset where a darker color shade of the name of the microbiota signifies a stronger (positive resp. negative) feature influence (Generalized JS Kernel).

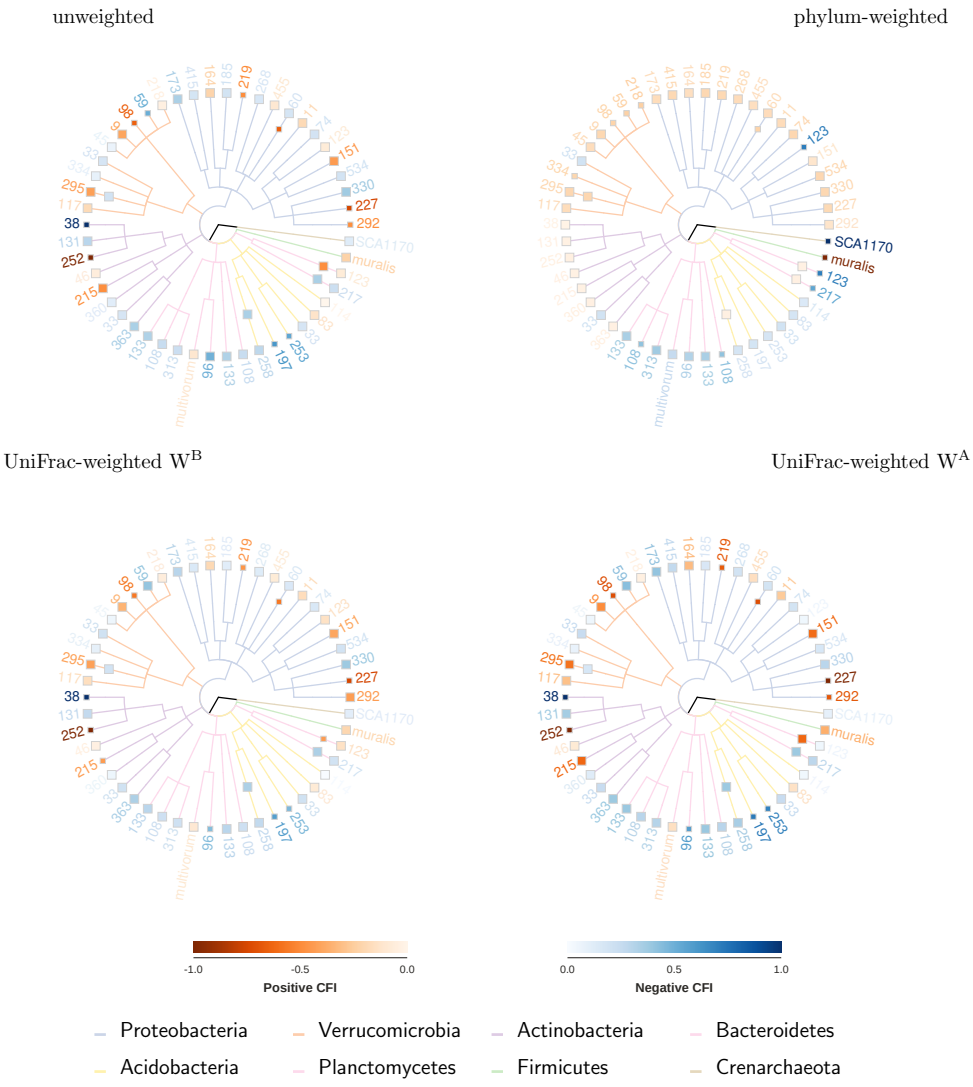



Figure 2.C.8: Scaled CFI values for centralparksoil dataset where a darker color shade of the name of the microbiota signifies a stronger (positive resp. negative) feature influence (Aitchison Kernel).

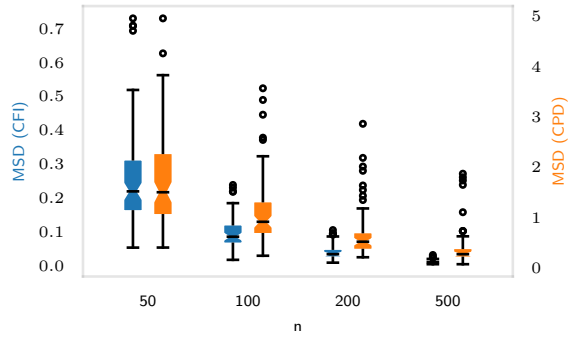



Figure 2.D.1: MSD of estimated CFI and CPD using `KernelBiome` estimator based on 100 random datasets for each sample size. For CPD, we calculate the true and estimated CPD based on 100 evenly spaced grid points within the range of $[0.001, 0.999]$ and the reported MSD is the average MSD over the 9 components. As the sample size $n$ increases the CFI and CPD estimates based on `KernelBiome` converge to the true population quantities.

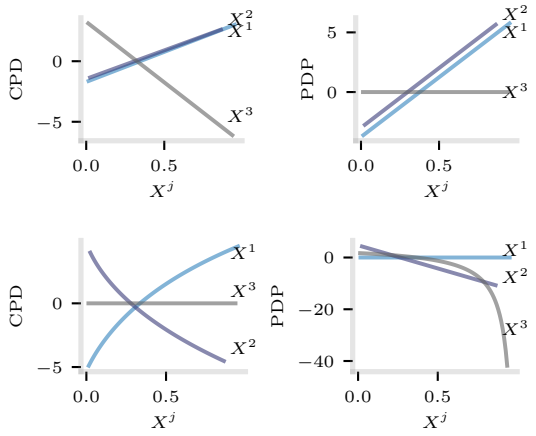



Figure 2.D.2: Top row: CPD and PDP plot based on $f_1$. Bottom row: CPD and PDP plot based on $f_2$. CPD reflects the true feature importance on the simplex while PDP does not.

## 2.E. Background on kernels

### 2.E.1 Connection between metrics and kernels

**Definition 2.E.1** (Metric, semi-metric, and quasi-metric)**.** A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *metric* if it satisfies

(a) $d(x, x) = 0$,

(b) $d(x, y) = d(y, x) \geq 0$,

(c) $d(x, y) \leq d(x, z) + d(y, z)$,

(d) $d(x, y) = 0 \Rightarrow x = y$.

It is called a *semi-metric* if it satisfies (a)-(c), and a *quasi-metric* if it satisfies (a)-(b). ♣

**Definition 2.E.2** (Function of negative type and Hilbertian metric)**.** A quasi-metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called of *negative-type* if for all $n \in \mathbb{N}$, all $x_1, \cdots, x_n \in \mathcal{X}$, and all $c_1, \cdots, c_n \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$, it holds that

$$\sum_{i,j=1}^{n} c_i c_j d^2(x_i, x_j) \leq 0. \tag{2.E.1}$$

If $d$ is a (semi-)metric, then $d$ is also called *Hilbertian.* ♣

**Theorem 2.E.3** (Sufficient and necessary conditions for isometric embeddings)**.** *A quasi-metric space $(X, d)$ can be isometrically embedded in a Hilbert space if and only if $d$ is of negative type.*

*Proof.* See [Wells and Williams, 2012, Theorem 2.4]. □

**Definition 2.E.4** ((conditionally) positive definite kernels)**.** A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (i.e., $\forall x, y \in \mathcal{X}, k(x, y) = k(y, x)$) is called a *positive definite kernel* if and only if for all $n \in \mathbb{N}$, all $x_1, \cdots, x_n \in \mathcal{X}$, and all $c_1, \cdots, c_n \in \mathbb{R}$, it holds that

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{2.E.2}$$

It is called a *conditional positive definite kernel* if instead of for all $c_1, \cdots, c_n \in \mathbb{R}$ condition (2.E.2) only holds for all $c_1, \cdots, c_n \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$. ♣

**Lemma 2.E.5.** *Let $\mathcal{X}$ be a non-empty set, fix $x_0 \in \mathcal{X}$ and let $k, \tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric functions satisfying for all $x, y \in \mathcal{X}$ that*

$$k(x, y) = \tilde{k}(x, y) - \tilde{k}(x, x_0) - \tilde{k}(y, x_0) + \tilde{k}(x_0, x_0) \tag{2.E.3}$$

*Then $k$ is positive definite if and only if $\tilde{k}$ is conditionally positive definite.*

*Proof.* Fix $n \in \mathbb{N}$, $c_1, \cdots, c_n \in \mathbb{R}$, and $x_0, x_1, \cdots, x_n \in \mathcal{X}$. Let $c_0 = -\sum_{i=1}^n c_i$, then we have

$$\begin{aligned}
\sum_{i,j=0}^{n} c_i c_j \tilde{k}(x_i, x_j) &= \sum_{i,j=1}^{n} c_i c_j \tilde{k}(x_i, x_j) + \sum_{i=1}^{n} c_i c_0 \tilde{k}(x_i, x_0) \\
&\quad + \sum_{j=1}^{n} c_0 c_j \tilde{k}(x_j, x_0) + c_0 c_0 \tilde{k}(x_0, x_0) \\
&= \sum_{i,j=1}^{n} c_i c_j \tilde{k}(x_i, x_j) - \sum_{i,j=1}^{n} c_i c_j \tilde{k}(x_i, x_0) \\
&\quad - \sum_{i,j=1}^{n} c_i c_j \tilde{k}(x_j, x_0) + \sum_{i,j=1}^{n} c_i c_j \tilde{k}(x_0, x_0) \\
&= \sum_{i,j=1}^{n} c_i c_j [\tilde{k}(x, y) - \tilde{k}(x, x_0) - \tilde{k}(y, x_0) + \tilde{k}(x_0, x_0)] \\
&= \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j).
\end{aligned} \tag{2.E.4}$$

Now, if $\tilde{k}$ is conditionally positive definite, then (2.E.4) implies that $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$, so $k$ is positive definite; if $k$ is positive definite, (2.E.4) implies that $\sum_{i,j=0}^n c_i c_j \tilde{k}(x_i, x_j \geq 0$ so $\tilde{k}$ is conditionally positive definite. This completes the proof of Lemma 2.E.5. □

**Lemma 2.E.6** (Shifted conditionally positive definite)**.** *Let $\mathcal{X}$ be a non-empty set and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel, then*

$$\tilde{k}(x, y) = k(x, y) + f(x) + f(y)$$

*is a conditionally positive definite kernel for all $f : \mathcal{X} \to \mathbb{R}$.*

*Proof.* The proof follows the exact same argument as the proof of Lemma 2.E.5. □

**Theorem 2.E.7** (Connection between Hilbertian semi-metrics and positive definite kernels)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be functions. If $k$ is a positive definite kernel and $d$ satisfies $d^2(x, y) = k(x, x) + k(y, y) - 2k(x, y)$, then $d$ is a Hilbertian semi-metric. On the other hand, for any $x_0 \in \mathcal{X}$, if $d$ is a Hilbertian semi-metric and and $k$ satisfies $k(x, y) = -\frac{1}{2}d^2(x, y) + \frac{1}{2}d^2(x, x_0) + \frac{1}{2}d^2(x_0, y)$, then $k$ is a pd kernel.*

The result is due to Schoenberg [1938].

*Proof.* We start with the first part. Assume that $k$ is a positive definite kernel and $d$ satisfies $d^2(x, y) = k(x, x) + k(y, y) - 2k(x, y)$. Then, $d$ is indeed a semi-metric by the following arguments:

(a) $d(x, x) = \sqrt{k(x, x) + k(x, x) - 2k(x, x)} = 0$,

(b) $d(x,y) = d(y,x)$, and since $k$ is positive definite, let $c_1 = 1$, $c_2 = -1$, $x_1 = x$, and $x_2 = y$,

$$\begin{aligned} 0 \leq \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) &= k(x_1, x_1) - k(x_1, x_2) - k(x_2, x_1) + k(x_2, x_2) \\ &= k(x,x) + k(y,y) - 2k(x,y) \\ &= d(x,y) \end{aligned}$$

(c) Since $k$ is a positive definite kernel, there exists a feature map $\varphi_k$ from $\mathcal{X}$ to an RKHS $\mathcal{H}_k$, and we have

$$\begin{aligned} ||\varphi_k(x) - \varphi_k(y)||^2_{\mathcal{H}_k} &= \langle \varphi_k(x) - \varphi_k(y), \varphi_k(x) - \varphi_k(y) \rangle_{\mathcal{H}_k} \\ &= \langle \varphi_k(x), \varphi_k(x) \rangle_{\mathcal{H}_k} + \langle \varphi_k(y), \varphi_k(y) \rangle_{\mathcal{H}_k} - 2\langle \varphi_k(x), \varphi_k(y) \rangle_{\mathcal{H}_k} \\ &= k(x,x) + k(y,y) - 2k(x,y) \\ &= d^2(x,y) \end{aligned}$$

Therefore, $d(x,z) \leq d(x,y) + d(y,z)$ follows from the triangle inequality of a norm.

To show $d$ is also Hilbertian, take any $n \in \mathbb{N}$, any $x_1, \cdots, x_n \in \mathcal{X}$, and any $c_1, \cdots, c_n \in \mathbb{R}$, we have

$$\begin{aligned} \sum_{i,j=1}^{n} c_i c_j d(x_i, x_j) &= \sum_{i=1}^{n} c_i k(x_i, x_i) \sum_{j=1}^{n} c_j + \sum_{j=1}^{n} c_i k(x_j, x_j) \sum_{i=1}^{n} c_i \\ &\quad - 2 \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \\ &= -2 \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \leq 0 \quad \text{(since } k \text{ is positive definite).} \end{aligned}$$

This proves the first part of the theorem.

For the second part, assume that $d$ is a Hilbertian semi-metric and $k$ satisfies $k(x,y) = -\frac{1}{2}d^2(x,y) + \frac{1}{2}d^2(x,x_0) + \frac{1}{2}d^2(x_0,y)$. Then, since $d$ is Hilbertian, $-d^2$ satisfies the requirement of a conditionally positive definite kernel (with the additional property that $-d^2(x,x) = 0$). Hence, by Lemma 2.E.5, $k$ is indeed positive definite. This completes the proof of Theorem 2.E.7. □

## 2.E.2 Dimensionality reduction and visualization with kernels

One important benefit of using the kernel approach is that we can leverage the kernels for dimensionality reduction and visualization, so that one can identify outliers in the data and further investigate them. In this section, we provide a short introduction on how to use kernels for multi-dimensional scaling and connect it to kernel PCA [Schölkopf et al., 2002].

Kernel methods project the compositional data into a (potentially) high-dimensional RKHS $\mathcal{H}_k$, which we now want to project into the low dimensional Euclidean space $\mathbb{R}^\ell$ (with $\ell \ll p$) such that the lower dimensional representation preserves information that helps separate the observations of different traits in the RKHS. That is, given observations $x_1, \cdots, x_n \in \mathbb{S}^{p-1}$ and a kernel $k$, we would like to define a map $\Phi : \mathcal{H}_k \to \mathbb{R}^\ell$ such that

$$\sum_{i,j=1}^{n} \|\langle k(x_i,\cdot), k(x_j,\cdot)\rangle_{\mathcal{H}_k} - \langle \Phi(k(x_i,\cdot)), \Phi(k(x_j,\cdot))\rangle_{\mathbb{R}^\ell}\|^2$$

is minimized. In matrix notation, this corresponds to solving

$$\arg\min_{Z\in\mathbb{R}^{n\times\ell}} \|K - ZZ^\top\|^2,$$

where the rows of $Z$ are $z_i = \Phi(k(x_i,\cdot)) \in \mathbb{R}^\ell$ for all $i \in \{1\cdots, n\}$ and $K \in \mathbb{R}^{n\times n}$ is the kernel Gram-matrix. This is similar to the classical multidimensional scaling (MDS) but measuring the similarity in the RKHS instead of in Euclidean space. By the Eckart-Young theorem [Eckart and Young, 1936], this minimization problem can be solved via the eigendecomposition of the matrix $K = V\Sigma V^\top$, and the optimal solution is

$$Z_{\text{opt}} = (V_1, \ldots, V_\ell)(\Sigma_{:\ell})^{\frac{1}{2}},$$

where $V_1, \ldots, V_\ell$ are the first $\ell$ columns of $V$ and $\Sigma_{:\ell}$ is the upper-left $(\ell \times \ell)$-submatrix of $\Sigma$. The optimal projection $\Phi_{\text{opt}}$ is then given for all $f \in \mathcal{H}_k$ by

$$\Phi_{\text{opt}}(f) = (\Sigma_{:\ell})^{-\frac{1}{2}}(V_1, \ldots, V_\ell)^\top \begin{pmatrix} \langle f, k(x_1,\cdot)\rangle_{\mathcal{H}_k} \\ \vdots \\ \langle f, k(x_n,\cdot)\rangle_{\mathcal{H}_k} \end{pmatrix}. \tag{2.E.5}$$

This in particular allows to project a new observations $w \in \mathbb{S}^{p-1}$ with the same projection that is $w \mapsto \Phi_{\text{opt}}(k(w,\cdot))$.

The projection in (2.E.5) depends on the origin of the RKHS $\mathcal{H}_k$. To remove this dependence, it may therefore be desirable to consider a centered version of the optimal projection. This can be achieved by considering the RKHS $\tilde{\mathcal{H}}_k$ consisting of the functions $\tilde{f}(\cdot) = f(\cdot) - \frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$ with $f \in \mathcal{H}_k$. To compute the optimal centered projection (2.E.5) for the RKHS $\tilde{\mathcal{H}}_k$, we only need to perform double centering on the kernel matrix $K$, i.e., $\tilde{K} = HKH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and replace $k(x,\cdot)$ by $\tilde{k}(x,\cdot) = k(x,\cdot) - \frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$. With the centering step, this procedure is equivalent to kernel PCA [Schölkopf et al., 2002]. The steps to obtain the lower-dimensional representation in matrix form are given in Algorithm 1.

**Algorithm 1:** Dimensionality reduction with kernels

**Input:** Training data $X_1, \ldots, X_n \in \mathbb{S}^{p-1}$, visualization data $X_1^{\text{new}}, \ldots, X_m^{\text{new}} \in \mathbb{S}^{p-1}$ (can be same as training data), kernel function $k$, dimension $l \in \{1, \ldots, p\}$, indicator whether to use centering $CenterK \in \{\texttt{True}, \texttt{False}\}$

**Function** `CenterKernelMatrix(`$K$, $\widetilde{K}$`)`**:**
  $K^{\text{center}} \leftarrow \widetilde{K} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top K - \frac{1}{n}\widetilde{K}\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top K\mathbf{1}\mathbf{1}^\top$
  **return** $K^{\text{center}}$

**for** $i, j \in \{1, \ldots, n\}$ **do**
  $K_{ij} \leftarrow k(X_i, X_j)$

**for** $i \in \{1, \ldots, m\}$ *and* $j \in \{1, \ldots, n\}$ **do**
  $K_{ij}^{\text{new}} \leftarrow k(X_i^{\text{new}}, X_j)$

**if** *CenterK* **then**
  $K^{\text{new}} \leftarrow$ `CenterKernelMatrix(`$K$, $K^{new}$`)`
  $K \leftarrow$ `CenterKernelMatrix(`$K$, $K$`)`

$V, \Sigma \leftarrow$ eigen decomposition of $K$
$Z \leftarrow K^{\text{new}}(V_1, \ldots, V_l)(\Sigma_{:l})^{-1/2}$

**Output:** $l$-dimensional representation $Z = (Z_1, \ldots, Z_m)^\top \in \mathbb{R}^{m \times l}$

### 2.E.3 Compositionally adjusted coordinate-wise contribution to each principle component

Given the optimal projection function $\Phi_{\text{opt}}$, define the function $F : \mathbb{S}^{p-1} \to \mathbb{R}^\ell$ for all $x \in \mathbb{S}^{p-1}$ by $F(x) = \Phi_{\text{opt}}(k(x, \cdot))$. We then call the components $F^1, \dots, F^\ell$ the principle components. Our goal is now to understand how each principle component is affected by changes in the different components of its arguments. For this, fix a principle component $r \in \{1, \dots, \ell\}$ and consider for each $j \in \{1, \dots, p\}$ the quantities

$$\mathbb{E}[F^r(\psi_j(X, c)) - F^r(X)],$$

where $c \in (0, 1)$ and $\psi_j$ the perturbation defined in Appendix 2.A. This is very similar in spirit as the CFI but with the derivative replaced by a difference and measures how much a perturbation of size $c$ in the $j$-th component effects the value of the $r$-th principle component. It is easily estimated by

$$\frac{1}{n} \sum_{i=1}^{n} F^r(\psi_j(X_i, c)) - F^r(X_i).$$

## 2.F. Proofs

### 2.F.1 Proof of Proposition 2.1

*Proof.* We start with the CFI. Fix $j \in \{1, \dots, p\}$ and $x \in \mathbb{S}^{p-1}$, then we can compute the derivative using the chain rule and the explicit form of the perturbation $\psi$ as follows

$$\begin{aligned}
\tfrac{d}{dc} f(\psi_j(x, c)) &= \left\langle \nabla f(\psi_j(x, c)), \tfrac{d}{dc} \psi_j(x, c) \right\rangle \\
&= \left\langle \nabla f(\psi_j(x, c)), \tfrac{d}{dc} s_c(x^1, \cdots, x^{j-1}, cx^j, x^{j+1}, \cdots, x^p)^\top \right\rangle \\
&= \left\langle \nabla f(\psi_j(x, c)), \tfrac{d}{dc} \frac{1}{\sum_{\ell \neq j}^p x^\ell + cx^j} (x^1, \cdots, x^{j-1}, cx^j, x^{j+1}, \cdots, x^p)^\top \right\rangle \\
&= \Bigg\langle \nabla f(\psi_j(x, c)), \\
&\qquad \frac{-x^j}{(\sum_{\ell \neq j}^p x^\ell + cx^j)^2} \left(x^1, \cdots, x^{j-1}, cx^j x^j - x^j(\textstyle\sum_{\ell \neq j}^p x^\ell + cx^j), x^{j+1}, \cdots, x^p\right)^\top \Bigg\rangle.
\end{aligned}$$

Evaluating, the derivative at $c = 1$ leads to

$$\tfrac{d}{dc} f(\psi_j(x, c))|_{c=1} = \left\langle \nabla f(x), x^j(e_j - x) \right\rangle, \tag{2.F.1}$$

where we used that $\psi_j(x, 1) = x$. Moreover, the gradient of $f$ in the case of the log-contrast model is given by

$$\nabla f(x) = \left( \frac{\beta_1}{x^1}, \dots, \frac{\beta_p}{x^p} \right)^\top. \tag{2.F.2}$$

Combining (2.F.1) and (2.F.2) together with the constraint $\sum_{k=1}^{p} \beta_k = 0$ implies that

$$\tfrac{d}{dc} f(\psi_j(x,c))|_{c=1} = -x^j \sum_{k \neq j} \beta_k + \beta^j (1 - x^j) = \beta_j.$$

Hence, taking the expectation leads to

$$I_j^j = \mathbb{E}[\tfrac{d}{dc} f(\psi_j(X,c))|_{c=1}] = \beta_j,$$

which proves the first part of the proposition.

Next, we show the result for the CPD. Fix $j \in \{1, \ldots, p\}$ and $z \in [0,1]$. Then $S_f^j(z)$ for the log-contrast model can be computed as follows

$$\begin{aligned}
S_f^j(z) &= \mathbb{E}[f(\varphi_j(X,z))] - \mathbb{E}[f(X)] \\
&= \sum_{\ell=1}^{p} \beta_\ell \mathbb{E}[\log(\varphi_j(X,z)^\ell)] - \mathbb{E}[f(X)] \\
&= \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(sX^\ell)] + \beta_j \log(z) - \mathbb{E}[f(X)] \\
&= \beta_j \log(z) + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(s)] + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \mathbb{E}[f(X)],
\end{aligned}$$

where $s = (1-z)/(\sum_{\ell \neq j}^{p} X^\ell)$. Using $\beta^j = -\sum_{\ell \neq j}^{p} \beta_\ell$ (which follows from the log-contrast model constraint on $\beta$) we can simplify this further and get

$$\begin{aligned}
S_f^j(z) &= \beta_j \log(z) + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(1-z)] - \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(\textstyle\sum_{k \neq j}^{p} X^k)] + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \mathbb{E}[f(X)] \\
&= \beta_j \log(z) - \beta_j \mathbb{E}[\log(1-z)] + \beta_j \mathbb{E}[\log(\textstyle\sum_{k \neq j}^{p} X^k)] + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \mathbb{E}[f(X)] \\
&= \beta_j \log\left(\frac{z}{1-z}\right) + \beta_j \mathbb{E}[\log(\textstyle\sum_{k \neq j}^{p} X^k)] + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \sum_{\ell=1}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] \\
&= \beta_j \log\left(\frac{z}{1-z}\right) + c,
\end{aligned}$$

with $c = \beta_j \mathbb{E}[\log(\sum_{k \neq j}^{p} X^k)] + \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \sum_{\ell=1}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)]$. Finally, assume $\beta^j = 0$, then it holds that

$$c = \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] - \sum_{\ell \neq j}^{p} \beta_\ell \mathbb{E}[\log(X^\ell)] = 0.$$

This completes the proof of Proposition 2.1. □

### 2.F.2 Proof of Theorem 2.1

*Proof.* We first prove (i). To see this, we apply the triangle inequality to get that

$$|\hat{I}^j_{\hat{f}_n} - I^j_{f^*}| \leq \underbrace{|\hat{I}^j_{\hat{f}_n} - \hat{I}^j_{f^*}|}_{=:A_n} + \underbrace{|\hat{I}^j_{f^*} - I^j_{f^*}|}_{=:B_n}. \tag{2.F.3}$$

Next, we consider the two terms $A_n$ and $B_n$ separately. We begin with $A_n$, by using the definition of the CFI together with (2.F.1) from the proof of Proposition 2.1. This leads to

$$\begin{aligned}
A_n &= \left| \frac{1}{n} \sum_{i=1}^n \left( \tfrac{d}{dc} \hat{f}_n(\psi(X_i, c)|_{c=1} - \tfrac{d}{dc} f^*(\psi(X_i, c)|_{c=1} \right) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \left\langle \nabla \hat{f}_n(X_i) - \nabla f^*(X_i), X_i^j (e_j - X_i) \right\rangle \right| \\
&\leq \frac{1}{n} \sum_{i=1}^n \left| \left\langle \nabla \hat{f}_n(X_i) - \nabla f^*(X_i), X_i^j (e_j - X_i) \right\rangle \right| \\
&\leq \frac{1}{n} \sum_{i=1}^n \left\| \nabla \hat{f}_n(X_i) - \nabla f^*(X_i) \right\|_2 \left\| X_i^j (e_j - X_i) \right\|_2 \\
&\leq \frac{1}{n} \sum_{i=1}^n \left\| \nabla \hat{f}_n(X_i) - \nabla f^*(X_i) \right\|_2,
\end{aligned}$$

where for the last three steps we used the triangle inequality, the Cauchy-Schwartz inequality and that $\|X_i^j(e_j - X_i)\|_2 \leq 1$ since $X_i \in \mathbb{S}^{p-1}$, respectively. By assumption, it therefore holds that $A_n \to 0$ in probability as $n \to \infty$. For the $B_n$ term, observe that using the same bounds it holds that

$$\mathbb{E}\left[ \left( \tfrac{d}{dc} f^*(\psi(X_i, c)|_{c=1} \right)^2 \right] = \mathbb{E}\left[ \left( \left\langle \nabla f^*(X_i), X_i^j (e_j - X_i) \right\rangle \right)^2 \right] \leq \mathbb{E}\left[ \|\nabla f^*(X_i)\|_2^2 \right].$$

By assumption that $\mathbb{E}\left[ \|\nabla f^*(X_i)\|_2^2 \right] < \infty$ this implies we can apply the weak law of large numbers to get for $n \to \infty$ that

$$\hat{I}^j_{f^*} = \frac{1}{n} \sum_{i=1}^n \tfrac{d}{dc} f^*(\psi(X_i, c)|_{c=1} \xrightarrow{P} \mathbb{E}\left[ \tfrac{d}{dc} f^*(\psi(X_i, c)|_{c=1} \right] = I^j_{f^*}.$$

This immediately implies that $B_n \to 0$ in probability as $n \to \infty$. Combining the convergence of $A_n$ and $B_n$ in (2.F.3) completes the proof of (i).

Next, we prove (ii). Fix $j \in \{1, \ldots, p\}$ and $z \in [0, 1]$ such that $z/(1-z) \in \operatorname{supp}(X^j / \sum_{\ell \neq j} X^\ell)$. By the definition of the perturbation $\varphi_j$ we get that

$$\varphi_j(X, z) = s(X^1, \cdots, X^{j-1}, \tfrac{z}{1-z} \textstyle\sum_{\ell \neq j} X^\ell, X^{j+1}, \cdots, X^p) \tag{2.F.4}$$

where $s = (1-z)/(\sum_{\ell\neq j}^{p} X^\ell)$. Next, using the assumption that $\mathrm{supp}(X) = \{x \in \mathbb{S}^{p-1} \mid x = w/(\sum_j w^j)$ with $w \in \mathrm{supp}(X^1) \times \cdots \times \mathrm{supp}(X^p)\}$ and that $z/(1-z) \in \mathrm{supp}(X^j/\sum_{\ell\neq j} X^\ell)$ we get that

$$\varphi_j(X, z) \in \mathrm{supp}(X^j) \tag{2.F.5}$$

almost surely.

By the triangle inequality it holds that

$$|\hat{S}^j_{\hat{f}_n}(z) - S^j_{f^*}(z)| \leq \underbrace{|\hat{S}^j_{\hat{f}_n}(z) - \hat{S}^j_{f^*}(z)|}_{=:C_n} + \underbrace{|\hat{S}^j_{f^*}(z) - S^j_{f^*}(z)|}_{=:D_n}. \tag{2.F.6}$$

We now consider the two terms $C_n$ and $D_n$ separately. First, we apply the triangle inequality to bound the $C_n$ term as follows.

$$\begin{aligned} C_n &= \left| \frac{1}{n}\sum_{i=1}^{n} (\hat{f}_n(\varphi_j(X_i, z)) - f^*(\varphi_j(X_i, z))) + \frac{1}{n}\sum_{i=1}^{n} (\hat{f}_n(X_i) - f^*(X_i)) \right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left| \hat{f}_n(\varphi_j(X_i, z)) - f^*(\varphi_j(X_i, z)) \right| + \frac{1}{n}\sum_{i=1}^{n} \left| \hat{f}_n(X_i) - f^*(X_i) \right| \\ &\leq 2 \sup_{x \in \mathrm{supp}(X)} \left| \hat{f}_n(x) - f^*(x) \right|, \end{aligned}$$

where for the last step we used a supremum bound together with (2.F.4). Hence, using the assumption that $\sup_{x\in\mathrm{supp}(X)} |\hat{f}_n(x) - f^*(x)| \xrightarrow{P} 0$ as $n \to \infty$, we get that $C_n \to \infty$ in probability as $n \to \infty$. Similarly, for the $D_n$ term we get that

$$\begin{aligned} D_n &= \left| \frac{1}{n}\sum_{i=1}^{n} f^*(\varphi_j(X_i, z)) - \mathbb{E}[f^*(\varphi_j(X_i, z))] + \frac{1}{n}\sum_{i=1}^{n} f^*(X_i) - \mathbb{E}[f^*(X_i)] \right| \\ &\leq \left| \frac{1}{n}\sum_{i=1}^{n} f^*(\varphi_j(X_i, z)) - \mathbb{E}[f^*(\varphi_j(X_i, z))] \right| + \left| \frac{1}{n}\sum_{i=1}^{n} f^*(X_i) - \mathbb{E}[f^*(X_i)] \right|. \end{aligned}$$

Since the $X_1, \ldots, X_n$ and hence $\varphi_j(X_1, z), \ldots, \varphi_j(X_n, z)$ are i.i.d. and bounded we can apply the weak law of large numbers to get that $D_n \to 0$ in probability as $n \to \infty$.

Finally, combining the convergence of $C_n$ and $D_n$ with (2.F.6) proves (ii) and hence completes the proof of Theorem 2.1. □

### 2.F.3 Proof of Proposition 2.2

*Proof.* For this proof, we denote by $\mathbb{S}^{p-1}$ the open instead of the closed simplex.

First, since $k_W$ is a positive definite kernel (see Section 2.1 in Appendix 2.B for a proof), it holds that the RKHS $\mathcal{H}_{k_W}$ can be expressed as the closure of

$$\mathcal{F} := \left\{ f : \mathbb{S}^{p-1} \times \mathbb{S}^{p-1} \to \mathbb{R} \,\middle|\, \exists n \in \mathbb{N}, z_1, \ldots, z_n \in \mathbb{S}^{p-1}, \alpha_1, \ldots, \alpha_n \in \mathbb{R} : f(\cdot) = \textstyle\sum_{i=1}^{n} \alpha_i k_W(z_i, \cdot) \right\}.$$

We now show that any function in $\mathcal{F}$ has the expression given in the statement of the proposition. Let $f \in \mathcal{F}$ be arbitrary with the expansion

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k_W(z_i, \cdot).$$

Then, for all $x \in \mathbb{S}^{p-1}$ it holds that

$$\begin{aligned}
f(x) &= \sum_{i=1}^{n} \alpha_i \sum_{j,\ell=1}^{p} W_{\ell,j} \log\left(\frac{z_i^\ell}{g(z_i)}\right) \log\left(\frac{x^j}{g(x)}\right) \\
&= \sum_{j=1}^{p} \left( \sum_{\ell=1}^{p} W_{\ell,j} \sum_{i=1}^{n} \alpha_i \log\left(\frac{z_i^\ell}{g(z_i)}\right)\right) \log\left(\frac{x^j}{g(x)}\right) \\
&= \sum_{j=1}^{p} \left( \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell \right) \log\left(\frac{x^j}{g(x)}\right) \\
&= \sum_{j=1}^{p} \left( \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell \right) \log\left(x^j\right) - \left( \sum_{j,\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell \right) \log\left(g(x)\right) \\
&= \sum_{j=1}^{p} \left( \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell \right) \log\left(x^j\right) - \left( \sum_{\ell=1}^{p} \widetilde{\beta}_\ell \right) \log\left(g(x)\right), \qquad (2.F.7)
\end{aligned}$$

where in the third line we defined $\widetilde{\beta}_\ell := \sum_{i=1}^{n} \alpha_i \log\left(\frac{z_i^\ell}{g(z_i)}\right)$ and in the last equation we used that $\sum_{j=1}^{p} W_{\ell,j} = 1$ for all $\ell \in \{1, \ldots, p\}$ by construction of $W$. Furthermore, we get that

$$\sum_{j=1}^{p} \widetilde{\beta}_j = \sum_{i=1}^{n} \alpha_i \left( \sum_{j=1}^{p} \log\left(z_i^j\right) - p \log(g(z_i)) \right) = \sum_{i=1}^{n} \alpha_i \left( \sum_{j=1}^{p} \log\left(z_i^j\right) - \sum_{j=1}^{p} \log\left(z_i^j\right) \right) = 0. \qquad (2.F.8)$$

Now, combining this with (2.F.7) and setting $\beta_j := \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell$ implies that

$$f(x) = \beta^\top \log(x),$$

where $\beta$ does not depend on $x$.

It remains to show that $\beta$ satisfies (i) $\sum_{j=1}^{p} \beta_j = 0$ and (ii) for all $\ell \in \{1, \ldots, m\}$ it holds for all $i, j \in P_\ell$ that $\beta_i = \beta_j$. For (i), we can use (2.F.8) and directly compute

$$\sum_{j=1}^{p} \beta_j = \sum_{j=1}^{p} \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell = \sum_{\ell=1}^{p} \widetilde{\beta}_\ell = 0.$$

Finally for (ii), fix $k \in \{1, \dots, m\}$ and $i, j \in P_k$, then it holds that

$$\begin{aligned}
\beta_j &= \sum_{\ell=1}^{p} W_{\ell,j} \widetilde{\beta}_\ell \\
&= \sum_{\ell=1}^{p} \sum_{r=1}^{m} \frac{1}{|P_r|} \mathbb{1}_{\{\ell, j \in P_r\}} \widetilde{\beta}_\ell \\
&= \sum_{\ell=1}^{p} \frac{1}{|P_k|} \widetilde{\beta}_\ell \\
&= \sum_{\ell=1}^{p} \sum_{r=1}^{m} \frac{1}{|P_r|} \mathbb{1}_{\{\ell, i \in P_r\}} \widetilde{\beta}_\ell \\
&= \sum_{\ell=1}^{p} W_{\ell,i} \widetilde{\beta}_\ell \\
&= \beta_i.
\end{aligned}$$

This completes the proof of Proposition 2.2. □

## 2.G. List of kernels implemented in `KernelBiome`

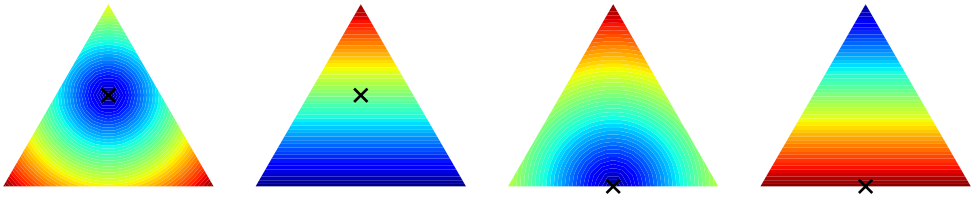
### 2.G.1 List of unweighted kernels

In this section we summarize the all kernels implemented in `KernelBiome` and visualize the metrics and kernels via heatmaps when $p = 3$. The reference points are the neutral point $u = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, a vertex $v = (1, 0, 0)$, a midpoint on a boundary $m = (\frac{1}{2}, \frac{1}{2}, 0)$, and an interior point $z = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$ of the simplex. For kernels we omit the neutral point, since $k(x, u) = 0$ for any $x \in \mathbb{S}^2$, as we centered our kernels at $u$.

**Linear metric & kernel**

$$d^2(x, y) = \sum_{j=1}^{p} (x^j - y^j)^2$$

$$k(x, y) = \Big( \sum_{j=1}^{p} x^j y^j \Big) - \tfrac{1}{p}$$

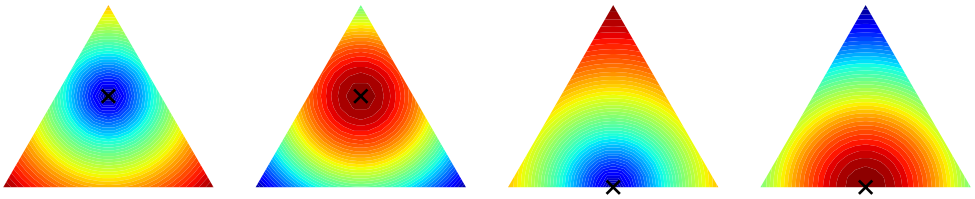


$d(x, z)$ $k(x, z)$ $d(x, m)$ $k(x, m)$

### RBF metric & kernel

$$d^2(x,y) = 2 - 2\exp\Big(\sum_{j=1}^{p} \frac{(x^j - y^j)^2}{2\sigma^2}\Big)$$

$$k(x,y) = \exp\Big(-\sum_{j=1}^{p} \frac{(x^j - y^j)^2}{2\sigma^2}\Big)$$

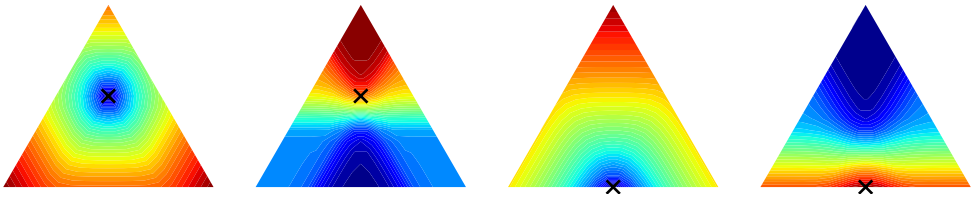


$d_{\sigma=\frac{1}{\sqrt{2}}}(x,z)$ $k_{\sigma=\frac{1}{\sqrt{2}}}(x,z)$ $d_{\sigma=\frac{1}{\sqrt{2}}}(x,m)$ $k_{\sigma=\frac{1}{\sqrt{2}}}(x,m)$

### Generalized-JS metric & kernel ($a < \infty, b \in [0.5, a)$)

$$d^2(x,y) = \frac{ab}{a-b}\sum_{j=1}^{p} \frac{2^{\frac{1}{b}}\Big[(x^j)^a + (y^j)^a\Big]^{\frac{1}{a}} - 2^{\frac{1}{a}}\Big[(x^j)^b + (y^j)^b\Big]^{\frac{1}{b}}}{2^{\frac{1}{a}+\frac{1}{b}}}$$

$$\begin{aligned} k(x,y) = -\frac{ab}{a-b}\cdot 2^{-(1+\frac{1}{a}+\frac{1}{b})}\sum_{j=1}^{p}\Bigg\{ & 2^{\frac{1}{b}}\Big(\Big[(x^j)^a + (y^j)^a\Big]^{\frac{1}{a}} - \Big[(x^j)^a + (\tfrac{1}{p})^a\Big]^{\frac{1}{a}} \\ & - \Big[(\tfrac{1}{p})^a + (y^j)^a\Big]^{\frac{1}{a}}\Big) - 2^{\frac{1}{a}}\Big(\Big[(x^j)^b + (y^j)^b\Big]^{\frac{1}{b}} - \Big[(x^j)^b + (\tfrac{1}{p})^b\Big]^{\frac{1}{b}} \\ & - \Big[(\tfrac{1}{p})^b + (y^j)^b\Big]^{\frac{1}{b}}\Big)\Bigg\} \end{aligned}$$

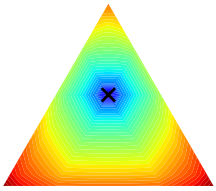


$d_{a=10,b=1}(x,z)$ $k_{a=10,b=1}(x,z)$ $d_{a=10,b=1}(x,z)$ $k_{a=10,b=1}(x,z)$

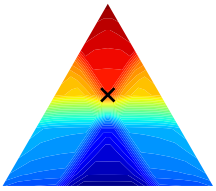
**Generalized-JS metric & kernel** ($a \to \infty, b < \infty$)

$$d^2(x,y) = \sum_{j=1}^{p} b\left\{ 2^{\frac{1}{b}} \cdot \max\{x^j, y^j\} - \left[ (x^j)^b + (y^j)^b \right]^{\frac{1}{b}} \right\}$$

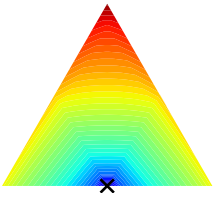
$$k(x,y) = -\frac{b}{2} \sum_{j=1}^{p} \left\{ 2^{\frac{1}{b}} \left( \max\{x^j, y^j\} - \max\{x^j, \tfrac{1}{p}\} - \max\{y^j, \tfrac{1}{p}\} \right) \right.$$
$$\left. - \left[ (x^j)^b + (y^j)^b \right]^{\frac{1}{b}} + \left[ (x^j)^b + (\tfrac{1}{p})^b \right]^{\frac{1}{b}} + \left[ (y^j)^b + (\tfrac{1}{p})^b \right]^{\frac{1}{b}} \right\}$$



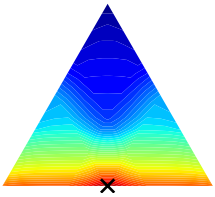
$d_{a=\infty,b=0.5}(x,z)$



$k_{a=\infty,b=0.5}(x,z)$



$d_{a=\infty,b=0.5}(x,m)$

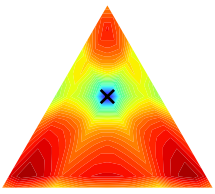


$k_{a=\infty,b=0.5}(x,m)$

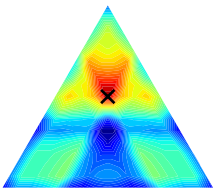
**Generalized-JS metric & kernel** ($a < \infty, b \to a$)

$$d^2(x,y) = \sum_{j=1}^{p} \left[\frac{(x^j)^b + (y^j)^b}{2}\right]^{\frac{1}{b}} \cdot \left[\frac{(x^j)^b}{(x^j)^b + (y^j)^b} \cdot \log \frac{2(x^j)^b}{(x^j)^b + (y^j)^b} + \frac{(y^j)^b}{(x^j)^b + (y^j)^b} \cdot \log \frac{2(y^j)^b}{(x^j)^b + (y^j)^b}\right]$$

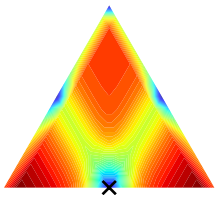
$$\begin{aligned} k(x,y) = -\frac{1}{2^{\frac{1}{b}+1}} \sum_{j=1}^{p} \Bigg\{ & \left[(x^j)^b + (y^j)^b\right]^{\frac{1}{b}-1} \cdot \Big((x^j)^b \cdot \log\left[2(x^j)^b\right] + (y^j)^b \log\left[2(y^j)^b\right] \\ & - \left[(x^j)^b + (y^j)^b\right] \cdot \log\left[(x^j)^b + (y^j)^b\right]\Big) \\ & - \left[(x^j)^b + \tfrac{1}{p^b}\right]^{\frac{1}{b}-1} \cdot \Big( - \left[(x^j)^b + (\tfrac{1}{p^b})\right] \cdot \log\left[(x^j)^b + \tfrac{1}{p^b}\right] \\ & + (x^j)^b \cdot \log\left[2(x^j)^b\right] + \tfrac{1}{p^b} \cdot \log\left[\tfrac{2}{p^b}\right]\Big) \\ & - \left[(y^j)^b + \tfrac{1}{p^b}\right]^{\frac{1}{b}-1} \cdot \Big( - \left[(y^j)^b + (\tfrac{1}{p^b})\right] \cdot \log\left[(y^j)^b + \tfrac{1}{p^b}\right] \\ & + (y^j)^b \cdot \log\left[2(y^j)^b\right] + \tfrac{1}{p^b} \cdot \log\left[\tfrac{2}{p^b}\right]\Big) \Bigg\} \end{aligned}$$



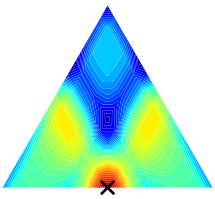
$d_{a=b=10}(x,z)$



$k_{a=b=10}(x,z)$



$d_{a=b=10}(x,m)$

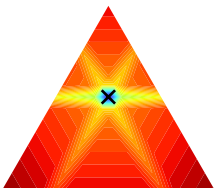


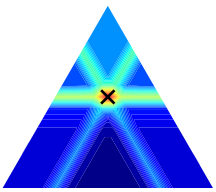
$k_{a=b=10}(x,m)$

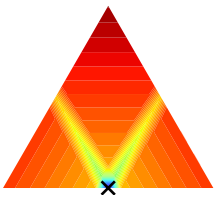
**Generalized-JS metric & kernel** ($a = b \to \infty$)

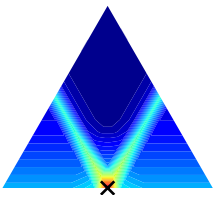
$$d^2(x,y) = \sum_{j=1}^{p} \max\{x^j, y^j\} \cdot \Big[\log(2)\mathbb{1}\{x^j \neq y^j\}\Big]$$

$$k(x,y) = -\frac{\log(2)}{2} \cdot \sum_{j=1}^{p} \Bigg\{ \max\{x^j, y^j\} \cdot \mathbb{1}\{x^j \neq y^j\} - \max\{x^j, \tfrac{1}{p}\} \cdot \mathbb{1}\{x^j \neq \tfrac{1}{p}\} - \max\{y^j, \tfrac{1}{p}\} \cdot \mathbb{1}\{y^j \neq \tfrac{1}{p}\} \Bigg\}$$


$d_{a=b=\infty}(x,z)$


$k_{a=b=\infty}(x,z)$


$d_{a=b=\infty}(x,m)$

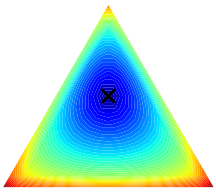

$k_{a=b=\infty}(x,m)$

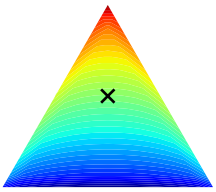
**Special Case: Hellinger - Generalized-JS metric & kernel** ($a = 1, b = \frac{1}{2}$)

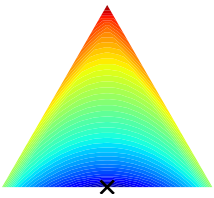
$$d^2(x,y) = \frac{\sqrt{2}}{2} \sum_{j=1}^{p} \left(\sqrt{x^j} - \sqrt{y^j}\right)^2$$

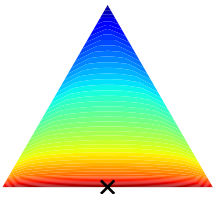
$$k(x,y) = \frac{\sqrt{2}}{4} + \frac{\sqrt{2}}{4} \sum_{j=1}^{p} \left\{ \sqrt{x^j y^j} - \frac{\sqrt{x^j} + \sqrt{y^j}}{\sqrt{p}} \right\}$$

This corresponds to $\frac{\sqrt{2}}{2}$ times the **Hellinger** metric and kernel.


$d_{a=1,b=0.5}(x,z)$


$k_{a=1,b=0.5}(x,z)$


$d_{a=1,b=0.5}(x,m)$

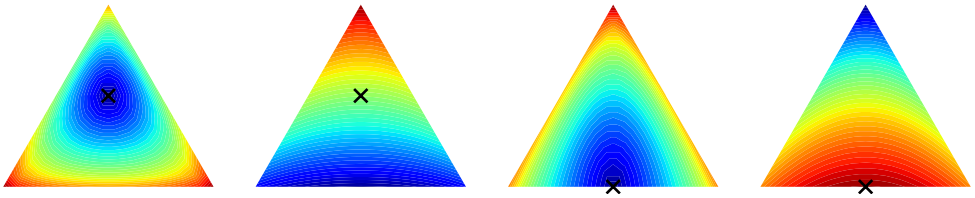

$k_{a=1,b=0.5}(x,m)$

**Special Case: Jenson-Shannon - Generalized-JS metric & kernel** ($a = 1, b = 1$)

$$d^2(x,y) = \frac{1}{2}\sum_{j=1}^{p} x^j \log\frac{2x^j}{x^j+y^j} + y^j \log\frac{2y^j}{x^j+y^j}$$

$$k(x,y) = -\frac{1}{4}\sum_{j=1}^{p}\left\{x^j \log\frac{x^j+\frac{1}{p}}{x^j+y^j} + y^j \log\frac{y^j+\frac{1}{p}}{x^j+y^j} - \frac{1}{p}\log\frac{4}{p^2(x^j+\frac{1}{p})(y^j+\frac{1}{p})}\right\}$$

This corresponds to the **Jenson-Shannon** metric and kernel.



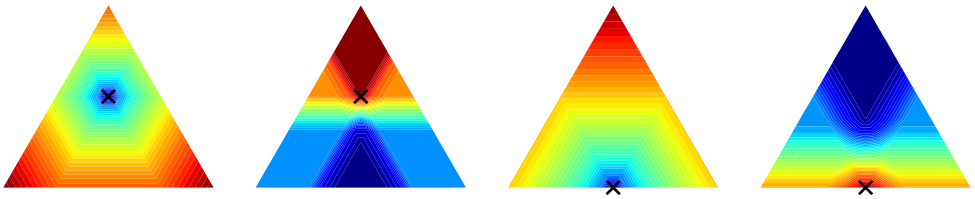
$d_{a=b=1}(x,z)$ $k_{a=b=1}(x,z)$ $d_{a=b=1}(x,z)$ $k_{a=b=1}(x,z)$

**Special Case: Total Variation - Generalized-JS metric & kernel** ($a = \infty, b = 1$)

$$d^2(x,y) = \sum_{j=1}^{p} |x^j - y^j|$$

$$k(x,y) = -\frac{1}{2}\sum_{j=1}^{p}\left\{|x^j - y^j| - |x^j - \tfrac{1}{p}| - |y^j - \tfrac{1}{p}|\right\}$$

This corresponds to 2 times the **total variation** metric and kernel.

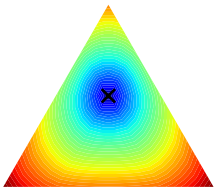


$d_{a=\infty,b=1}(x,z)$ $k_{a=\infty,b=1}(x,z)$ $d_{a=\infty,b=1}(x,z)$ $k_{a=\infty,b=1}(x,z)$

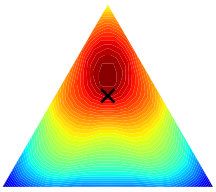
**Hilbertian metric & kernel** ($a < \infty, b > -\infty$)

$$d^2(x,y) = \sum_{j=1}^{p} \frac{2^{\frac{1}{b}}\left[(x^j)^a + (y^j)^a\right]^{\frac{1}{a}} - 2^{\frac{1}{a}}\left[(x^j)^b + (y^j)^b\right]^{\frac{1}{b}}}{2^{\frac{1}{a}} - 2^{\frac{1}{b}}}$$

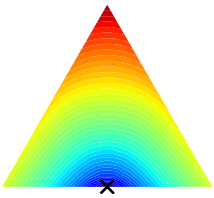
$$\begin{aligned} k(x,y) = -\frac{1}{2(2^{\frac{1}{a}} - 2^{\frac{1}{b}})} \sum_{j=1}^{p} \Bigg\{ & 2^{\frac{1}{b}} \Big( \left[(x^j)^a + (y^j)^a\right]^{\frac{1}{a}} - \left[(x^j)^a + (\tfrac{1}{p})^a\right]^{\frac{1}{a}} \\ & - \left[(y^j)^a + (\tfrac{1}{p})^a\right]^{\frac{1}{a}} \Big) - 2^{\frac{1}{a}} \Big( \left[(x^j)^b + (y^j)^b\right]^{\frac{1}{b}} - \left[(x^j)^b + (\tfrac{1}{p})^b\right]^{\frac{1}{b}} \\ & - \left[(y^j)^b + (\tfrac{1}{p})^b\right]^{\frac{1}{b}} \Big) \Bigg\} \end{aligned}$$



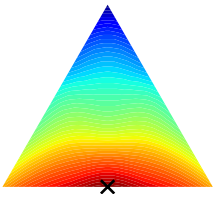
$d_{a=10,b=-1}(x,z)$



$k_{a=10,b=-1}(x,z)$



$d_{a=10,b=-1}(x,m)$

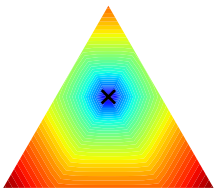


$k_{a=10,b=-1}(x,m)$

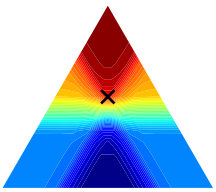
**Hilbertian metric & kernel** ($a \to \infty, b > -\infty$)

$$d^2(x,y) = \sum_{j=1}^{p} b \left\{ 2^{\frac{1}{b}} \cdot \max\{x^j, y^j\} - \left[(x^j)^b + (y^j)^b\right]^{\frac{1}{b}} \right\}$$

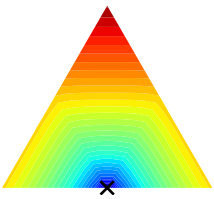
$$\begin{aligned} k(x,y) = -\frac{1}{2(1 - 2^{\frac{1}{b}})} \sum_{j=1}^{p} \Bigg\{ & 2^{\frac{1}{b}} \Big( \max\{x^j, y^j\} - \max\{x^j, \tfrac{1}{p}\} - \max\{y^j, \tfrac{1}{p}\} \Big) \\ & - \left[(x^j)^b + (y^j)^b\right]^{\frac{1}{b}} + \left[(x^j)^b + (\tfrac{1}{p})^b\right]^{\frac{1}{b}} + \left[(y^j)^b + (\tfrac{1}{p})^b\right]^{\frac{1}{\beta}} \Bigg\} \end{aligned}$$



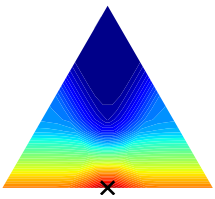
$d_{a=\infty,b=-10}(x,z)$



$k_{a=\infty,b=-10}(x,z)$



$d_{a=\infty,b=-10}(x,m)$

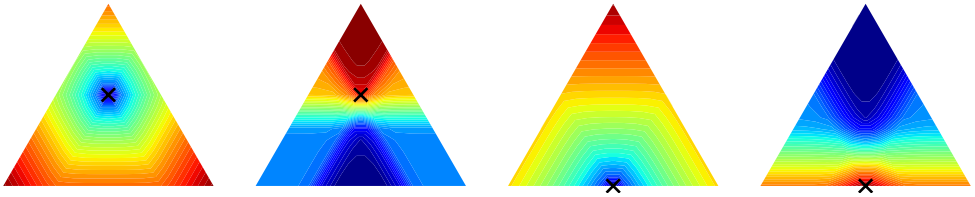


$k_{a=\infty,b=-10}(x,m)$

**Hilbertian metric & kernel** ($a < \infty, b \to -\infty$)

$$d^2(x,y) = \frac{1}{2^{\frac{1}{a}} - 1} \sum_{j=1}^{p} \left\{ \left[ (x^j)^a + (y^j)^a \right]^{\frac{1}{a}} - 2^{\frac{1}{a}} \cdot \min\{x^j, y^j\} \right\}$$

$$k(x,y) = -\frac{1}{2(2^{\frac{1}{a}} - 1)} \sum_{j=1}^{p} \left\{ \left[ (x^j)^a + (y^j)^a \right]^{\frac{1}{a}} - \left[ (x^j)^a + (\tfrac{1}{p})^a \right]^{\frac{1}{a}} - \left[ (y^j)^a + (\tfrac{1}{p})^a \right]^{\frac{1}{a}} - 2^{\frac{1}{a}} \left[ \min\{x^j, y^j\} - \min\{x^j, \tfrac{1}{p}\} - \min\{y^j, \tfrac{1}{p}\} \right] \right\}$$



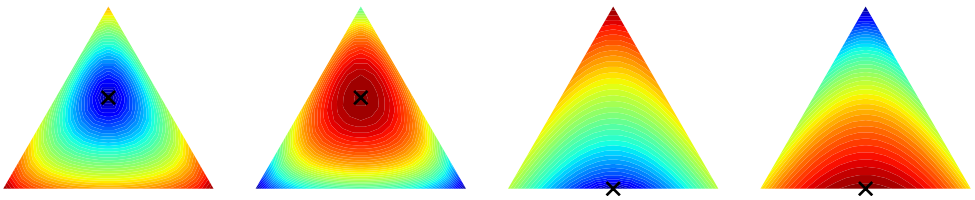
$d_{a=10,b=-\infty}(x,z)$ $k_{a=10,b=-\infty}(x,z)$ $d_{a=10,b=-\infty}(x,m)$ $k_{a=10,b=-\infty}(x,m)$

**Special Case: Chi-square - Hilbertian metric & kernel** ($a = 1, b = -1$)

$$d^2(x,y) = \frac{1}{3} \sum_{j=1}^{p} \frac{(x^j - y^j)^2}{x^j + y^j}$$

$$k(x,y) = -\frac{1}{6} \sum_{j=1}^{p} \left\{ \frac{(x^j - y^j)^2}{x^j + y^j} - \frac{(x^j - \frac{1}{p})^2}{x^j + \frac{1}{p}} - \frac{(y^j - \frac{1}{p})^2}{y^j + \frac{1}{p}} \right\}$$

This corresponds to $\frac{1}{3}$ of the **chi-square** metric and kernel.



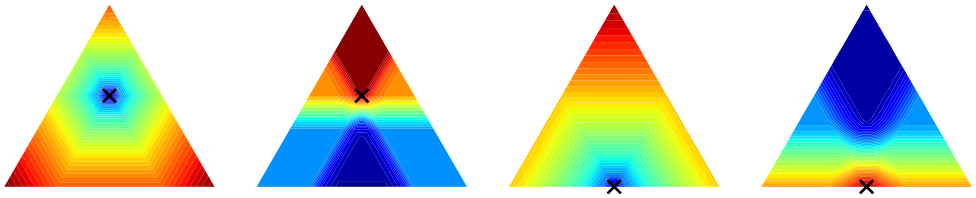
$d_{a=1,b=-1}(x,z)$ $k_{a=1,b=-1}(x,z)$ $d_{a=1,b=-1}(x,m)$ $k_{a=1,b=-1}(x,m)$

**Special Case: Total Variation - Hilbertian metric & kernel** ($a = 1, b = -\infty$)

$$d^2(x,y) = \frac{1}{2}\sum_{j=1}^{p} |x^j - y^j|$$

$$k(x,y) = -\frac{1}{4}\sum_{j=1}^{p}\left\{|x^j - y^j| - |x^j - \tfrac{1}{p}| - |y^j - \tfrac{1}{p}|\right\}$$

This corresponds to the **total variation** metric and kernel.



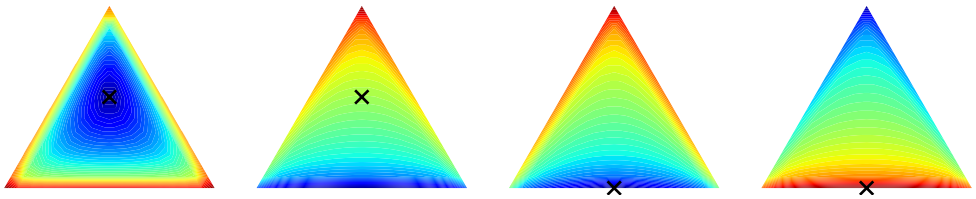
$d_{a=1,b=-\infty}(x,z)$ $k_{a=1,b=-\infty}(x,z)$ $d_{a=1,b=-\infty}(x,m)$ $k_{a=1,b=-\infty}(x,m)$

**Aitchison metric & kernel**

$$d^2(x,y) = \sum_{j=1}^{p}\left(\log\frac{x^j + c}{g(x+c)} - \log\frac{y^j + c}{g(y+c)}\right)^2$$

$$k(x,y) = \sum_{j=1}^{p}\log\frac{x^j + c}{g(x+c)}\log\frac{y^j + c}{g(y+c)}$$

where $g(x) = \sqrt[p]{\prod_{j=1}^{p} x^j}$ is the geometric mean of $x$.



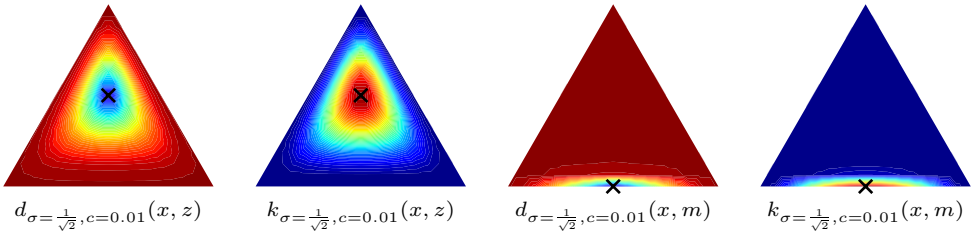
$d_{c=0.01}(x,z)$ $k_{c=0.01}(x,z)$ $d_{c=0.01}(x,m)$ $k_{c=0.01}(x,m)$

**Aitchison-RBF metric & kernel**

$$d^2(x,y) = 2 - 2\exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\Big[\log\frac{x^j+c}{g(x+c)} - \log\frac{y^j+c}{g(y+c)}\Big]^2\Big)$$

$$k(x,y) = \exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\Big[\log\frac{x^j+c}{g(x+c)} - \log\frac{y^j+c}{g(y+c)}\Big]^2\Big)$$

where $g(x) = \sqrt[p]{\prod_{j=1}^{p} x^j}$ is the geometric mean of $x$.

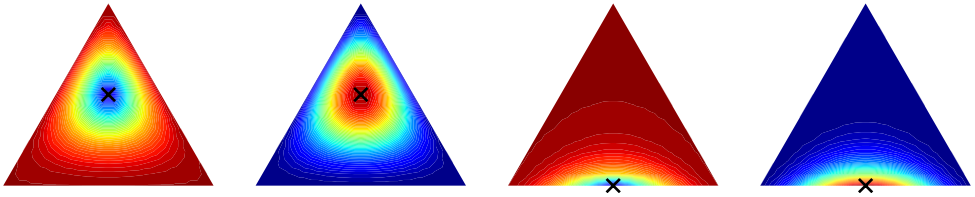


$d_{\sigma=\frac{1}{\sqrt{2}},c=0.01}(x,z)$ $k_{\sigma=\frac{1}{\sqrt{2}},c=0.01}(x,z)$ $d_{\sigma=\frac{1}{\sqrt{2}},c=0.01}(x,m)$ $k_{\sigma=\frac{1}{\sqrt{2}},c=0.01}(x,m)$

**Heat diffusion metric & kernel**

$$d^2(x,y) = 2\cdot(4\pi t)^{-\frac{p}{2}}\cdot\Big[1-\exp\Big(-\tfrac{1}{t}\arccos^2\big(\sum_{j=1}^{p}\sqrt{x^j y^j}\big)\Big)\Big]$$

$$k(x,y) = (4\pi t)^{-p/2}\cdot\exp\Big(-\tfrac{1}{t}\arccos^2\big(\sum_{j=1}^{p}\sqrt{x^j y^j}\big)\Big)$$



$d_{t=0.1}(x,z)$ $k_{t=0.1}(x,z)$ $d_{t=0.1}(x,m)$ $k_{t=0.1}(x,m)$

## 2.G.2 List of weighted kernels

As discussed in Appendix 2.B, all kernels can also be modified to include a weight matrix $W \in \mathbb{R}^{p\times p}$. Below, we list the explicit forms of all weighted kernels as they are implemented in `KernelBiome` package. As before, let $g(x) = \sqrt[p]{\prod_{j=1}^{p} x^j}$ be the geometric mean of $x$.

**Linear kernel**

$$k(x,y) = \sum_{j,\ell=1}^{p} W_{j,\ell}\Big(x^j y^\ell - \tfrac{x^j}{p} - \tfrac{y^\ell}{p} + \tfrac{1}{p^2}\Big)$$

**RBF kernel**

$$k(x,y) = \exp\Big(-\sum_{j,\ell=1}^{p} \frac{W_{j,\ell}(x^j - y^\ell)^2}{2\sigma^2}\Big)$$

**Generalized-JS kernel** ($a < \infty, b \in [0.5, a]$)

$$\begin{aligned} k(x,y) = -\frac{ab}{a-b}\cdot 2^{-(1+\frac{1}{a}+\frac{1}{b})} \sum_{j,\ell=1}^{p} W_{j,\ell}\Bigg\{ & 2^{\frac{1}{b}}\Big(\Big[(x^j)^a + (y^\ell)^a\Big]^{\frac{1}{a}} - \Big[(x^j)^a + (\tfrac{1}{p})^a\Big]^{\frac{1}{a}} \\ & - \Big[(\tfrac{1}{p})^a + (y^\ell)^a\Big]^{\frac{1}{a}}\Big) - 2^{\frac{1}{a}}\Big(\Big[(x^j)^b + (y^\ell)^b\Big]^{\frac{1}{b}} - \Big[(x^j)^b + (\tfrac{1}{p})^b\Big]^{\frac{1}{b}} \\ & - \Big[(\tfrac{1}{p})^b + (y^\ell)^b\Big]^{\frac{1}{b}}\Big)\Bigg\} \end{aligned}$$

**Generalized-JS kernel** ($a \to \infty, b < \infty$)

$$\begin{aligned} k(x,y) = -\frac{b}{2}\sum_{j,\ell=1}^{p} W_{j,\ell}\Bigg\{ & 2^{\frac{1}{b}}\Big(\max\{x^j, y^\ell\} - \max\{x^j, \tfrac{1}{p}\} - \max\{y^\ell, \tfrac{1}{p}\}\Big) \\ & - \Big[(x^j)^b + (y^\ell)^b\Big]^{\frac{1}{b}} + \Big[(x^j)^b + (\tfrac{1}{p})^b\Big]^{\frac{1}{b}} + \Big[(y^\ell)^b + (\tfrac{1}{p})^b\Big]^{\frac{1}{b}}\Bigg\} \end{aligned}$$

**Generalized-JS kernel** ($a < \infty, b \to a$)

$$\begin{aligned} k(x,y) = -\frac{1}{2^{\frac{1}{b}+1}}\sum_{j,\ell=1}^{p} W_{j,\ell}\Bigg\{ & \Big[(x^j)^b + (y^\ell)^b\Big]^{\frac{1}{b}-1}\cdot\Big((x^j)^b\cdot\log\Big[2(x^j)^b\Big] + (y^\ell)^b\log\Big[2(y^\ell)^b\Big] \\ & - \Big[(x^j)^b + (y^\ell)^b\Big]\cdot\log\Big[(x^j)^b + (y^\ell)^b\Big]\Big) \\ & - \Big[(x^j)^b + \tfrac{1}{p^b}\Big]^{\frac{1}{b}-1}\cdot\Big(-\Big[(x^j)^b + (\tfrac{1}{p^b})\Big]\cdot\log\Big[(x^j)^b + \tfrac{1}{p^b}\Big] \\ & + (x^j)^b\cdot\log\Big[2(x^j)^b\Big] + \tfrac{1}{p^b}\cdot\log\Big[\tfrac{2}{p^b}\Big]\Big) \\ & - \Big[(y^\ell)^b + \tfrac{1}{p^b}\Big]^{\frac{1}{b}-1}\cdot\Big(-\Big[(y^\ell)^b + (\tfrac{1}{p^b})\Big]\cdot\log\Big[(y^\ell)^b + \tfrac{1}{p^b}\Big] \\ & + (y^\ell)^b\cdot\log\Big[2(y^\ell)^b\Big] + \tfrac{1}{p^b}\cdot\log\Big[\tfrac{2}{p^b}\Big]\Big)\Bigg\} \end{aligned}$$

**Generalized-JS kernel** ($a = b \to \infty$)

$$k(x,y) = -\frac{\log(2)}{2} \cdot \sum_{j,\ell=1}^{p} W_{j,\ell} \Big\{ \max\{x^j, y^\ell\} \cdot \mathbb{1}\{x^j \neq y^\ell\} - \max\{x^j, \tfrac{1}{p}\} \cdot \mathbb{1}\{x^j \neq \tfrac{1}{p}\} - \max\{y^\ell, \tfrac{1}{p}\} \cdot \mathbb{1}\{y^\ell \neq \tfrac{1}{p}\} \Big\}$$

**Hilbertian kernel** ($a < \infty, b > -\infty$)

$$k(x,y) = -\frac{1}{2(2^{\frac{1}{a}} - 2^{\frac{1}{b}})} \sum_{j,\ell=1}^{p} W_{j,\ell} \Big\{ 2^{\frac{1}{b}} \Big( \Big[ (x^j)^a + (y^\ell)^a \Big]^{\frac{1}{a}} - \Big[ (x^j)^a + (\tfrac{1}{p})^a \Big]^{\frac{1}{a}} - \Big[ (y^\ell)^a + (\tfrac{1}{p})^a \Big]^{\frac{1}{a}} \Big) - 2^{\frac{1}{a}} \Big( \Big[ (x^j)^b + (y^\ell)^b \Big]^{\frac{1}{b}} - \Big[ (x^j)^b + (\tfrac{1}{p})^b \Big]^{\frac{1}{b}} - \Big[ (y^\ell)^b + (\tfrac{1}{p})^b \Big]^{\frac{1}{b}} \Big) \Big\}$$

**Hilbertian kernel** ($a \to \infty, b > -\infty$)

$$k(x,y) = -\frac{1}{2(1 - 2^{\frac{1}{b}})} \sum_{j,\ell=1}^{p} W_{j,\ell} \Big\{ 2^{\frac{1}{b}} \Big( \max\{x^j, y^\ell\} - \max\{x^j, \tfrac{1}{p}\} - \max\{y^\ell, \tfrac{1}{p}\} \Big) - \Big[ (x^j)^b + (y^\ell)^b \Big]^{\frac{1}{b}} + \Big[ (x^j)^b + (\tfrac{1}{p})^b \Big]^{\frac{1}{b}} + \Big[ (y^\ell)^b + (\tfrac{1}{p})^b \Big]^{\frac{1}{b}} \Big\}$$

**Hilbertian kernel** ($a < \infty, b \to -\infty$)

$$k(x,y) = -\frac{1}{2(2^{\frac{1}{a}} - 1)} \sum_{j,\ell=1}^{p} W_{j,\ell} \Big\{ \Big[ (x^j)^a + (y^\ell)^a \Big]^{\frac{1}{a}} - \Big[ (x^j)^a + (\tfrac{1}{p})^a \Big]^{\frac{1}{a}} - \Big[ (y^\ell)^a + (\tfrac{1}{p})^a \Big]^{\frac{1}{a}} - 2^{\frac{1}{a}} \Big[ \min\{x^j, y^\ell\} - \min\{x^j, \tfrac{1}{p}\} - \min\{y^\ell, \tfrac{1}{p}\} \Big] \Big\}$$

**Aitchison kernel**

$$k(x,y) = \sum_{j,\ell=1}^{p} W_{j,\ell} \log \frac{x^j + c}{g(x+c)} \log \frac{y^\ell + c}{g(y+c)}$$

**Aitchison RBF kernel**

$$k(x,y) = \exp\Big( -\frac{1}{2\sigma^2} \sum_{j,\ell=1}^{p} W_{j,\ell} \Big[ \log \frac{x^j + c}{g(x+c)} - \log \frac{y^\ell + c}{g(y+c)} \Big]^2 \Big)$$

**Heat diffusion kernel**

$$k(x,y) = (4\pi t)^{-p/2} \cdot \exp\Big( -\tfrac{1}{t} \arccos^2 \Big( \sum_{j,\ell=1}^{p} W_{j,\ell} \sqrt{x^j y^\ell} \Big) \Big)$$

# 3 Causal change point detection and localization

Shimeng Huang, Jonas Peters, Niklas Pfister

**Abstract**

Detecting and localizing change points in sequential data is of interest in many areas of application. Various notions of change points have been proposed, such as changes in mean, variance, or the linear regression coefficient. In this work, we consider settings in which a response variable $Y$ and a set of covariates $X = (X^1, \dots, X^{d+1})$ are observed over time and aim to find changes in the causal mechanism generating $Y$ from $X$. More specifically, we assume $Y$ depends linearly on a subset of the covariates and aim to determine at what time points either the dependency on the subset or the subset itself changes. We call these time points causal change points (CCPs) and show that they form a subset of the commonly studied regression change points. We propose general methodology to both detect and localize CCPs. Although motivated by causality, we define CCPs without referencing an underlying causal model. The proposed definition of CCPs exploits a notion of invariance, which is a purely observational quantity but – under additional assumptions – has a causal meaning. For CCP localization, we propose a loss function that can be combined with existing multiple change point algorithms to localize multiple CCPs efficiently. We evaluate and illustrate our methods on simulated datasets and two real datasets on Beijing air quality and Swiss monetary policy, respectively.

## 3.1 Introduction

Change point detection (i.e., testing the existence of change points) and localization (i.e., estimating the location of change points) have been of interest for several decades dating back to Page [1954, 1955]. We consider an offline setting where we have a sequence of independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ with covariates $X_i \in \mathbb{R}^{d+1}$ and a response $Y_i \in \mathbb{R}$. For all $i \in \{1, \dots, n\}$, denote by $\mathbb{P}_i^{X,Y}$ the joint distribution of $(X_i, Y_i)$, which may change across $i$. We call a time point $k \in \{2, \dots, n\}$ a *change point* if the joint distributions at time points $k$ and $k-1$ differ, that is, if $\mathbb{P}_k^{X,Y} \neq \mathbb{P}_{k-1}^{X,Y}$ [see also Brodsky and Darkhovsky, 1993]. Instead of considering general change points as defined above, one may consider a more restrictive definition of change points, e.g., time points

where there is a change in mean, variance or conditional distribution. Depending on the application at hand, certain types of change points may be more relevant than others. In many applications, the goal is to detect or localize changes in the relationship between the covariates and the response.

In economics and other fields, "structural changes", that is, changes in regression models, have been extensively studied over the last few decades. These include linear and nonlinear regression models, as well as non-parametric regression models. Under linear regression settings, Bai [1996] and Perron et al. [2020] propose tests for detecting changes in the regression parameter and the residual distribution; Hansen [2000] proposes a test that detects changes in the regression parameter while allowing for changes in the marginal distribution of the covariates; Bai [1997b] considers localizing one structural change while allowing lagged and trending covariates, and Bai [1997a] and Bai and Perron [1998, 2003] analyze the estimation of multiple change points. Andrews [1993] considers testing an unknown change point in part of the parameter vector in nonlinear regression models. Testing for changes in nonparametric regression models has been considered by, e.g., Orváth and Kokoszka [2002]. In recent years, structural changes in high-dimensional regression models have also been studied [e.g., Leonardi and Bühlmann, 2016, Wang et al., 2021a]. Reviews are provided by Aue and Horváth [2013], for example, who focus on methods detecting structural change that allow for serial dependence; Truong et al. [2020] consider algorithms that can be characterized by a cost function, a search method, and a constraint on the number of changes.

By definition, structural changes refer to changes in the conditional distribution of $Y$ given all covariates $X$. While in many applications such changes are useful, it may also be of interest to have some type of mechanistic understanding of the changes in order to assess their relevance. For example, if we assume there is an underlying causal model generating the distribution over the variables $(X, Y)$, then a structural change, that is, a change in the conditional distribution of $Y$ given $X$, can have different causal explanations: It could either indicate a change in causal relationship between $Y$ and $X$, or it could merely correspond to shifts in the distribution of $X$ that do not affect the causal dependence of $Y$ on $X$. The ability to distinguish between such changes can be useful in many applications as it allows practitioners to pay particular attention to the more fundamental changes. In this work, we characterize these changes based on reversing the idea of causal invariance — also known as autonomy or modularity [e.g., Haavelmo, 1943, Aldrich, 1989] — which gives the change points a causal interpretation under a causal model but is still meaningful otherwise. The idea of causal invariance has been used in invariant causal prediction proposed by Peters et al. [2016] and its sequential counterpart [Pfister et al., 2019] for discovering the causal predictors of a response variable, where the conditional distribution of the response variable given its causal predictors is assumed to be unchanged across environments (respectively, time). Our paper shows that this idea also proves useful when detecting and localizing change points. To our knowledge, detecting or localizing change points that can have a causal interpretation have not been studied with one exception on detecting local causal mechanism changes in directed acyclic graphs (DAGs) in Huang et al. [2020], where the definition of the causal mechanism is different from ours, specifically they assume that the parent set of

each node is fixed.

**Notation 3.1.1.** *We observe a sequence of independent observations* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *with covariates* $X_i \in \mathbb{R}^{d+1}$ *and a response* $Y_i \in \mathbb{R}$. *To avoid explicitly stating intercepts, we assume* $X_i^{d+1} = 1$ *for all* $i \in \{1, \ldots, n\}$, *and let* $\mathcal{S} := \big\{S \subseteq \{1, \ldots, d+1\} \mid d+1 \in S\big\}$. *We let* $\mathcal{I}$ *be the set of all subsets of* $\{1, \ldots, n\}$ *that are sequences of consecutive indices of length greater than or equal to* 2 *which we refer to as "intervals". For all* $S \in \mathcal{S}$ *we denote* $X_i^S \in \mathbb{R}^{|S|}$ *as the column vector of covariates* $(X_i^j)_{j \in S}$ *(sorted in ascending order of the indices). We denote by* $\mathbf{X} := (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$ *and* $\mathbf{Y} := (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^{n \times 1}$ *the design matrix of the covariates and the matrix of responses, respectively. For all* $I \in \mathcal{I}$, *we denote by* $\mathbf{X}_I$ *and* $\mathbf{Y}_I$ *the submatrices formed by the rows of* $\mathbf{X}$ *and* $\mathbf{Y}$ *indexed by* $I$, *respectively (sorted in ascending order of the indices), and additionally for all* $S \in \mathcal{S}$, $\mathbf{X}_I^S$ *denotes the submatrix of* $\mathbf{X}$ *formed by the rows indexed by* $I$ *and columns indexed by* $S$.

This paper is organized as follows. In Section 3.2, we define regression change points and causal change points. Section 3.3 focuses on the detection problem and introduces a simple procedure. In Section 3.4, we consider the localization problem and propose two different methods: one that tests candidates and one that minimizes a loss function. Numerical experiments and a real data application are given in Section 3.5.

## 3.2 Regression change points and causal change points

To distinguish between structural and causal changes, we first formally define the time points of structural changes below, which we call *regression change points*.

**Definition 3.2.1** (Regression change point (RCP))**.** For all $i \in \{1, \ldots, n\}$, assume that $\mathbb{E}[X_i X_i^\top]$ is invertible and define the population ordinary least squares (OLS) coefficient as $\beta_i^{\mathrm{OLS}} := \mathbb{E}[X_i X_i^\top]^{-1} \mathbb{E}[X_i Y_i]$ and the corresponding residual as $\epsilon_i := Y_i - X_i^\top \beta_i^{\mathrm{OLS}}$. Then, a time point $k \in \{2, \ldots, n\}$ is called a *regression change point* (RCP) if

$$\text{either} \quad \beta_k^{\mathrm{OLS}} \neq \beta_{k-1}^{\mathrm{OLS}} \quad \text{or} \quad \epsilon_k \overset{d}{\neq} \epsilon_{k-1}.$$

♣

While we do not assume that the conditional mean of $Y$ given $X$ is linear, the definition of RCPs implies that if $I \in \mathcal{I}$ is an interval without an RCP, then there exists a vector $\beta \in \mathbb{R}^{d+1}$ and a distribution $F_\epsilon$ such that for all $i \in I$ it holds that

$$Y_i = X_i^\top \beta + \epsilon_i \quad \text{and} \quad \mathbb{E}[X_i \epsilon_i] = 0,$$

with $\epsilon_i \sim F_\epsilon$ and $\beta_i^{\mathrm{OLS}} = \beta$.

RCPs characterize changes in the conditional mean model. However, even though these changes are sometimes interpreted as a proxy for a change in causality, it is well-known that this interpretation can be misleading. The following example illustrates this.

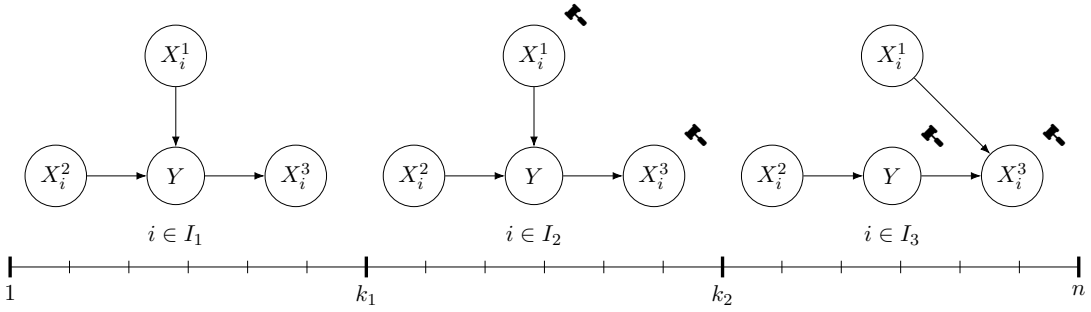



Figure 3.2.1: Illustration of the data generating model in Example 3.2.2 rolled out over time. The model remains fixed between 1 and $k_1 - 1$, between $k_1$ and $k_2 - 1$ and between $k_2$ and $n$. We intervene at two time points $k_1$ and $k_2$, and the hammers indicate on which node these interventions act with respect to the previous time interval. The population OLS coefficient $\beta_i^{\text{OLS}}$ changes at both time points $k_1$ and $k_2$ due to the interventions (see details in Appendix 3.A). However, only at $k_2$ the causal mechanism of $Y$ changes (at $k_1$ the causal mechanism of $Y$ remains unchanged).

**Example 3.2.2** (RCPs in linear SCMs)**.** Let $\{1, \dots, n\}$ be partitioned into three disjoint time intervals $I_1 = \{1, \dots, k_1 - 1\}$, $I_2 = \{k_1, \dots, k_2 - 1\}$ and $I_3 = \{k_2, \dots, n\}$. For all $i \in \{1, \dots, n\}$ consider the linear structural causal model (SCM), see also Section 3.2.1, over the variables $(X_i^1, X_i^2, X_i^3, X_i^4, Y_i)$ given by $X_i^4 := 1$ as the intercept and

$$X_i^S := A_i X_i + \alpha_i Y + \epsilon_i^X \tag{3.2.1a}$$

$$Y_i := \beta_i^\top X_i + \epsilon_i^Y, \tag{3.2.1b}$$

where $S = \{1, 2, 3\}$, $\epsilon_i^X = (\epsilon_i^{X^1}, \epsilon_i^{X^2}, \epsilon_i^{X^3})$ and $\epsilon_i^Y$ are jointly independent noise vectors with mean zero, $A_i \in \mathbb{R}^{3\times 4}$, $\beta_i \in \mathbb{R}^4$ and $\alpha_i \in \mathbb{R}^3$ are parameters such that the SCM induces the graphs in Figure 3.2.1 (the intercept variable $X_i^4$ is omitted). The specific values for the parameters $A_i$, $\beta_i$ and $\alpha_i$, as well as the variances of $\epsilon_i^X$ and $\epsilon_i^Y$ for $i \in \{1, \dots, n\}$ are given in Appendix 3.A.

In this example, at both time points $k_1$ and $k_2$, the joint distribution of $(X_i, Y_i)$ and in particular the population OLS parameter $\beta_i^{\text{OLS}}$ changes (i.e., $k_1$ and $k_2$ are both RCPs by Definition 3.2.1). However, the causal mechanism of the response $Y$ with respect to $X$ as specified in (3.2.1b) only changes at $k_2$. Our proposed notion of causal change points defined in Definition 3.2.4 below is able to capture this distinction. This example also highlights the invariance property of causal models: Interventions that do not act directly on the response $Y_i$ may change $Y_i \mid X_i$ but they keep $Y_i \mid X_i^{\text{PA}(Y_i)}$ invariant, where $\text{PA}(Y_i) \subseteq \{1, \dots, 4\}$ denotes the causal parents of $Y_i$. A more formal treatment of causal models is provided in Section 3.2.1.

♠

Example 3.2.2 shows that it is possible that the conditional expectation of $Y$ given $X$ can change even though the causal mechanism of how $Y$ is affected by $X$ remains

fixed. We propose to distinguish between changes only in the full conditional expectation of $Y$ given $X$ and changes that manifest in differences in the conditional expectations of $Y$ given $X^S$ for all $S \in \mathcal{S}$. Arguably, the second notion of change is of a more fundamental nature and indicates a more drastic shift in the data generating process. To formalize this notion which we call causal change points (see Definition 3.2.4 below) we first define the population OLS coefficient and the corresponding residuals based on subsets of covariates.

**Definition 3.2.3** (Population OLS given subsets of covariates)**.** Assume that $\mathbb{E}[X_i X_i^\top]$ is invertible for all $i \in \{1, \dots, n\}$. For all $S \in \mathcal{S}$ and all $i \in \{1, \dots, n\}$, the *population OLS coefficient given* $S$ is defined as $\beta_i^{\mathrm{OLS}}(S) \in \mathbb{R}^{d+1}$ satisfying

$$\beta_i^{\mathrm{OLS}}(S)^S = \mathbb{E}\left[X_i^S (X_i^S)^\top\right]^{-1} \mathbb{E}\left[X_i^S Y_i\right]$$

and $\beta_i^{\mathrm{OLS}}(S)^j = 0$ for all $j \in \{1, \dots, d+1\} \setminus S$. The corresponding *population OLS residual given* $S$ is defined as $\epsilon_i(S) := Y_i - X_i^\top \beta_i^{\mathrm{OLS}}(S)$. We use the convention that $\beta_i^{\mathrm{OLS}} = \beta_i^{\mathrm{OLS}}(\{1, \dots, d+1\})$ and $\epsilon_i = \epsilon_i(\{1, \dots, d+1\})$. ♣

Using this definition, we can now define what we call causal change points, the time points at which for all subsets of covariates $S \in \mathcal{S}$, either the population OLS coefficient given $S$ or the distribution of the population OLS residual given $S$ differs from previous time points (see Definition 3.2.4). Even though we call these changes "causal", the definition does not rely on an underlying causal model. Nevertheless, the definition of causal change points is motivated by the fact that under additional causal assumptions, they correspond to changes in the causal mechanism of $Y$ on $X$. We discuss this connection in Section 3.2.1.

**Definition 3.2.4** (Causal change point (CCP))**.** A time point $k \in \{2, \dots, n\}$ is called a *causal change point* (CCP) if for all $S \in \mathcal{S}$

$$\text{either} \quad \beta_k^{\mathrm{OLS}}(S) \neq \beta_{k-1}^{\mathrm{OLS}}(S) \quad \text{or} \quad \epsilon_k(S) \overset{d}{\neq} \epsilon_{k-1}(S).$$

♣

By definition, CCPs form a subset of RCPs. We refer to RCPs that are not CCPs as *non-causal change points* (NCCPs). An alternative way to characterize CCPs is via sets $S \in \mathcal{S}$ for which the population OLS coefficient and residual distribution given $S$ remain unchanged within a time interval.

**Definition 3.2.5** (Invariant set)**.** For a time interval $I \in \mathcal{I}$, a set $S \in \mathcal{S}$ is called an *$I$-invariant set* if there exists a vector $\beta \in \mathbb{R}^{d+1}$ and a distribution $F$ such that for all $i \in I$,

$$\beta_i^{\mathrm{OLS}}(S) = \beta \quad \text{and} \quad \epsilon_i(S) \overset{\mathrm{iid}}{\sim} F.$$

♣

The following proposition characterizes CCPs in terms of invariant sets. The proof, given in Appendix 3.D for completeness, follows directly from the definitions.

**Proposition 3.2.6** (Alternative characterization of CCP)**.** *A time point $k \in \{2, \dots, n\}$ is a CCP if and only if there does not exist a $\{k-1, k\}$-invariant set $S \in \mathcal{S}$.*

In the following section, we discuss how CCPs relate to changes of causal mechanism when assuming an underlying causal model.

### 3.2.1 Causal models as data generating models

We now formalize changes in causal mechanisms and relate them to CCPs. To this end, we introduce a class of SCMs [e.g., Pearl, 2009] that satisfy the assumptions of our sequential model, and discuss under which additional causal assumptions, CCPs correspond to causal mechanism changes. Furthermore, we argue in Example 3.2.8 that even if these assumptions are not satisfied, CCPs capture meaningful changes.

**Setting 1** (Sequential linear SCM with hidden variables)**.** *Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^{d+1} \times \mathbb{R}$ be a sequence of observed variables and $(H_1, \dots, H_n) \in \mathbb{R}^q$ a sequence of unobserved variables. For all $i \in \{1, \dots, n\}$ consider an SCM over $(H_i, X_i, Y_i)$ given by $X_i^{d+1} := 1$ as intercept and*

$$X_i^{S^\star} := A_i X_i + \alpha_i Y_i + h_i(H_i, \epsilon_i^X) \tag{3.2.2a}$$

$$Y_i := \beta_i^\top X_i + g_i(H_i, \epsilon_i^Y), \tag{3.2.2b}$$

*where $S^\star = \{1, \dots, d\}$, $H_i$, $\epsilon_i^X$ and $\epsilon_i^Y$ are jointly independent, $g_i$ and $h_i$ are arbitrary measurable functions such that $\mathbb{E}[h_i(H_i, \epsilon_i^Y)] = \mathbb{E}[g_i(H_i, \epsilon_i^Y)] = 0$. Furthermore, the parameters in (3.2.2a) and (3.2.2b) are such that for all $i \in \{1, \dots, n\}$ the induced graph[1] is directed and acyclic. For all $i \in \{2, \dots, n-1\}$ the set of (observed) parent variables of $Y_i$ is given by $PA(Y_i) = \{j \in \{1, \dots, d+1\} \mid \beta_i^j \neq 0\}$.*

Given such a causal model, we can characterize what CCPs correspond to under certain conditions. In Proposition 3.2.7, we show that as long as the noise term of $Y$ remains uncorrelated with its parents, a CCP indicates a change in either the causal coefficient $\beta_i$ or in the noise term $g_i(H_i, \epsilon_i^Y)$.

**Proposition 3.2.7.** *Assume Setting 1, let $k \in \{2, \dots, n\}$ be a fixed time point and assume that for all $i \in \{k-1, k\}$ it holds that*

$$\mathbb{E}[X_i^{PA(Y_i)} g_i(H_i, \epsilon_i^Y)] = 0. \tag{3.2.3}$$

*Then, it holds that*

$$k \text{ is a CCP} \implies \beta_k \neq \beta_{k-1} \text{ or } g_k(H_k, \epsilon_k^Y) \overset{d}{\neq} g_{k-1}(H_{k-1}, \epsilon_{k-1}^Y).$$

[1] For all time points $i \in \{1, \dots, n\}$ the graph is constructed by taking the observed variables $X_i^1, \dots, X_i^4, Y_i$ as nodes and adding a directed edge from node $V$ to $W$ if variable $V$ appears with a non-zero coefficient in the structural equation of variable $W$.

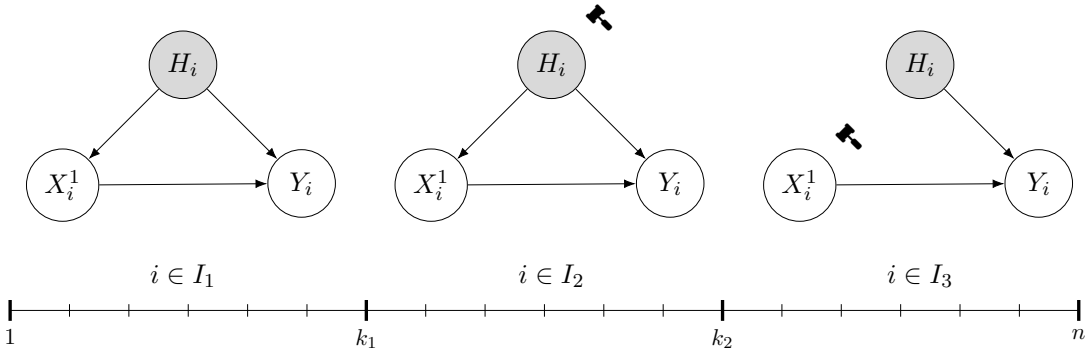



Figure 3.2.2: Illustration of the data generating model in Example 3.2.8 rolled out over time. The model remains fixed between 1 and $k_1 - 1$, between $k_1$ and $k_2 - 1$ and between $k_2$ and $n$. We intervene at two time points $k_1$ and $k_2$, and the hammers indicate on which node these interventions act with respect to the previous time interval. Even though both $k_1$ and $k_2$ are CCPs, the causal mechanism of $Y$ with respect to $X$ only changes at $k_1$ (the noise term $H_i + \epsilon_i^Y$ changes in distribution) but not at $k_2$ (neither the causal coefficient nor the noise term's distribution changes).

A proof is given in Appendix 3.D. In the following example, we illustrate that the statement of Proposition 3.2.7 may be false if there is hidden confounding in the sense that (3.2.3) is violated.

**Example 3.2.8** (CCPs with hidden confounding)**.** For all $i \in \{1, \dots, n\}$, consider the linear SCM over the variables $(H_i, X_i^1, X_i^2, Y_i)$ given by $X_i^2 := 1$ as the intercept and

$$X_i^1 := \alpha_i H_i + \epsilon_i^{X^1} \tag{3.2.4a}$$
$$Y_i := X_i^1 + H_i + \epsilon_i^Y, \tag{3.2.4b}$$

where $\epsilon_i^{X^1}, \epsilon_i^Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for all $i \in \{1, \dots, n\}$, $H_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for all $i \in \{1, \dots, k_1 - 1\}$, $H_i \overset{\text{iid}}{\sim} \mathcal{N}(0,2)$ for all $i \in \{k_1, \dots, n\}$, $\alpha_i = 1$ for all $i \in \{1, \dots, k_2 - 1\}$, and $\alpha_i = 0$ for all $i \in \{k_2, \dots, n\}$. For all $i \in \{1, \dots, n\}$, the variable $H_i$ is unobserved. Define the intervals $I_1 = \{1, \dots, k_1 - 1\}$, $I_2 = \{k_1, \dots, k_2 - 1\}$, and $I_3 = \{k_2, \dots, n\}$. The corresponding DAGs are shown in Figure 3.2.2.

Here, both $k_1$ and $k_2$ are CCPs. To see this, consider the population OLS parameter given $S_1 = \{1, 2\}$ which is equal to $\beta_i^{\text{OLS}}(S_1) = (c_i, 0)^\top$, where

$$c_i = \frac{\text{Cov}(X_i^1, Y_i)}{\mathbb{V}(X_i^1)} = \begin{cases} 3/2 & i \in I_1 \\ 5/3 & i \in I_2 \\ 1 & i \in I_3. \end{cases}$$

Hence, for all $k \in \{k_1, k_2\}$ the set $S_1$ is not $\{k-1, k\}$-invariant. Moreover, the population

OLS residual given $S_2 = \{2\}$ is given by

$$\epsilon_i(S_2) \sim \begin{cases} \mathcal{N}(0,6) & i \in I_1 \\ \mathcal{N}(0,10) & i \in I_2 \\ \mathcal{N}(0,4) & i \in I_3. \end{cases}$$

Again this implies that for all $k \in \{k_1, k_2\}$ the set $S_2$ is not $\{k-1, k\}$-invariant. By Proposition 3.2.6, both $k_1$ and $k_2$ are CCPs. This shows that in the case of hidden confounding between $Y$ and $X^{\mathrm{PA}(Y)}$, it is no longer true that the existence of a CCP implies a change in the causal mechanism of $Y$ (as specified in (3.2.4b)) as in the unconfounded case considered in Proposition 3.2.7. Nevertheless, we consider time points like $k_1$ and $k_2$ as conceptually different from other RCPs in that they represent changes in what can be thought of as the observed causal mechanism. ♠

## 3.3 Causal change point detection

We now consider how to detect CCPs, that is, given a time interval $I \in \mathcal{I}$, we would like to decide whether there exists a CCP $k \in I$. For a fixed time interval $I \in \mathcal{I}$, the absence of CCPs in $I$ is equivalent (see Proposition 3.2.6) to the null hypothesis

$$\mathcal{H}_0^I : \exists S \in \mathcal{S} \text{ s.t. } S \text{ is } I\text{-invariant.} \tag{3.3.5}$$

The goal is to construct a (possibly randomized) hypothesis test $\phi_I : \mathbb{R}^{|I| \times (d+1)} \times \mathbb{R}^{|I|} \to \{0,1\}$ for $\mathcal{H}_0^I$. The test $\phi_I$ is a function of the data $(\mathbf{X}_I, \mathbf{Y}_I)$ and rejects the null hypothesis $\mathcal{H}_0^I$ if $\phi_I(\mathbf{X}_I, \mathbf{Y}_I) = 1$ and does not reject it if $\phi_I(\mathbf{X}_I, \mathbf{Y}_I) = 0$. $\phi_I$ is said to be level $\alpha \in (0,1)$ if $\sup_{P \in \mathcal{H}_0^I} \mathbb{P}_P(\phi_I(\mathbf{X}_I, \mathbf{Y}_I) = 1) \leq \alpha$, and it is said to have power $\beta \in (0,1)$ against an alternative $P \notin \mathcal{H}_0^I$ if $\mathbb{P}_P(\phi_I(\mathbf{X}_I, \mathbf{Y}_I) = 1) = \beta$.

Testing $\mathcal{H}_0^I$ can be split up into a multiple testing problem by considering for all $S \in \mathcal{S}$ the null hypothesis

$$\mathcal{H}_{0,S}^I : S \text{ is } I\text{-invariant.} \tag{3.3.6}$$

This null hypothesis equals the hypothesis that the population OLS coefficient and residuals given $S$ do not change. For such settings, tests have been derived previously (see Section 3.3.1). Given a collection of tests $(\phi_I^S)_{S \in \mathcal{S}}$ for the null hypotheses $(\mathcal{H}_{0,S}^I)_{S \in \mathcal{S}}$ that are at level $\alpha$, we can combine them to a test for $\mathcal{H}_0^I$ as follows.

**Proposition 3.3.1.** *Let $(\phi_I^S)_{S \in \mathcal{S}}$ be a family of tests for the hypotheses $(\mathcal{H}_{0,S}^I)_{S \in \mathcal{S}}$ where for all $S \in \mathcal{S}$, $\phi_I^S : \mathbb{R}^{|I| \times |S|} \times \mathbb{R}^{|I|} \to \{0,1\}$, and $\phi_I^S$ is level $\alpha \in (0,1)$. Then the test $\phi_I : \mathbb{R}^{|I| \times (d+1)} \times \mathbb{R}^{|I|} \to \{0,1\}$ defined for all $x \in \mathbb{R}^{|I| \times (d+1)}$ and all $y \in \mathbb{R}^{|I|}$ by*

$$\phi_I(x,y) := \begin{cases} 1 & \text{if } \min_{S \in \mathcal{S}} \phi_I^S(x^S, y) = 1 \\ 0 & \text{otherwise} \end{cases}$$

*is level $\alpha$ for $\mathcal{H}_0^I$.*

The proof of this proposition is straightforward and is included in Appendix 3.D.

### 3.3.1 Tests for $\mathcal{H}_{0,S}^I$

We first introduce the following definition of population OLS parameter and residuals over a time interval given a subset of covariates.

**Definition 3.3.2** (Population OLS over an interval given a subset of covariates)**.** Let $I \in \mathcal{I}$ and assume $\sum_{i\in I} \mathbb{E}[X_i X_i^\top]$ is invertible. For all $S \in \mathcal{S}$, the *population OLS parameter given $S$ over $I$* is the vector $\beta_I^{\mathrm{OLS}}(S) \in \mathbb{R}^{d+1}$ where

$$\Big(\beta_I^{\mathrm{OLS}}(S)\Big)^S = \Big[\sum_{i\in I} \mathbb{E}[X_i^S (X_i^S)^\top]\Big]^{-1} \sum_{i\in I} \mathbb{E}(X_i^S Y_i)$$

and $\big(\beta_I^{\mathrm{OLS}}(S)\big)^j = 0$ for all $j \in \{1,\dots,d+1\} \setminus S$. For all $\ell \in \{1,\dots,n\}$, the *population OLS residual at $\ell$ given $S$ over $I$* is given by $\epsilon_\ell^I(S) := Y_\ell - X_\ell^\top \beta_I^{\mathrm{OLS}}(S)$ with the convention that $\epsilon^I(S) = \Big(Y_\ell - X_\ell^\top \beta_I^{\mathrm{OLS}}(S)\Big)_{\ell\in I} \in \mathbb{R}^{|I|}$. ♣

One way to test the null hypothesis $\mathcal{H}_{0,S}^I$ is to first divide the interval $I$ into two sub-intervals $I_1 := \{\min(I),\dots,\min(I) + \lfloor \frac{|I|}{2} \rfloor - 1\}$ and $I_2 := \{\min(I) + \lfloor \frac{|I|}{2} \rfloor,\dots,\max(I)\}$. Then, $\mathcal{H}_{0,S}^I$ in (3.3.5) implies

$$\mathcal{H}_{0,S}^{I_1,I_2} : \beta_{I_1}^{\mathrm{OLS}}(S) = \beta_{I_2}^{\mathrm{OLS}}(S) \text{ and } \epsilon^{I_1}(S) \overset{d}{=} \epsilon^{I_2}(S). \tag{3.3.7}$$

The reverse implication, however, is not true in general: $\mathcal{H}_{0,S}^{I_1,I_2}$ does not generally imply $\mathcal{H}_{0,S}^I$. This means that a test that is level for $\mathcal{H}_{0,S}^{I_1,I_2}$ is level for $\mathcal{H}_{0,S}^I$, however, power against some of the alternatives of $\mathcal{H}_{0,S}^I$ could be reduced. (3.3.7) can be tested by e.g., the Chow test [Chow, 1960]. Details of the Chow test are given in Appendix 3.F and we refer to applying the Chow test after splitting an interval into two as the *split Chow test*. A version of this test has also been suggested for Invariant Causal Prediction [Peters et al., 2016]. As an alternative, one can use the procedure proposed by Pfister et al. [2019]: instead of two sub-intervals, one considers a pre-defined grid over the time indices and combines test statistics computed based on resampling scaled versions of the residuals.

## 3.4 Causal change point localization

We now discuss two approaches for estimating the locations of CCPs. The first approach is based on testing candidates. By Definition 3.2.4, CCPs are a subset of RCPs. Thus, if we are given the set of RCPs, we can use the detection method described in Section 3.3 to identify the CCPs among them. An alternative approach is based on a loss function that aims to detect the CCPs directly. For localizing multiple CCPs, we can combine the proposed loss function (see Definition 3.4.4) with existing multiple change point localization algorithms. Popular multiple change point localization algorithms are often of two types: algorithms based on dynamic programming [e.g., Hawkins, 1976, Killick et al., 2012] and greedy algorithms [e.g., Vostrikova, 1981, Fryzlewicz, 2014].

In order to estimate the locations of all CCPs, both approaches rely on statistical methods for detecting changes in both regression parameters and the residual distributions.

Throughout this section, we assume there exists a set of $q \in \{1, \dots, n-2\}$ CCPs $\mathcal{T} \coloneqq \{\tau_1, \dots, \tau_q\}$, where $\tau_i < \tau_{i+1}$ for all $i \in \{1, \dots, q-1\}$ and we use the convention that $\tau_0 \coloneqq 1$ and $\tau_{q+1} \coloneqq n+1$.

### 3.4.1 Causal change point localization by pruning candidates

Assume we are given a candidate set $\mathcal{K} = \{k_1, \dots, k_l\} \subseteq \{2, \dots, n-1\}$ of potential CCPs. This could, for example, be the set of RCPs or a superset of the RCPs (see Definition 3.2.1). For the purpose of this section, we assume that the true CCPs are contained in $\mathcal{K}$ but in practice one would estimate the set $\mathcal{K}$ using existing methods for localizing RCPs [e.g., Bai, 1997a], which may lead to violations of this assumption. We can then prune the candidate set $\mathcal{K}$ by testing whether a candidate $k_j$ is indeed a causal change point considering the interval $I_j = \{k_{j-1}, \dots, k_{j+1} - 1\}$ (with the convention that $k_0 = 1$ and $k_{l+1} = n+1$) and using a test for $\mathcal{H}_0^{I_j}$ discussed in Section 3.3. The detailed procedure is provided in Algorithm 2 in Appendix 3.B.

Proposition 3.4.1 gives lower bounds (which are functions of the properties of the test) of the probability that Algorithm 2 localizes only the true CCPs and the probability that Algorithm 2 localizes all the true CCPs when the candidate set is a superset of the true CCPs.

**Proposition 3.4.1.** *Denote by $\mathcal{T} \subseteq \{2, \dots, n\}$ the set of CCPs and by $\mathcal{K} = \{k_1, \dots, k_L\} \subseteq \{2, \dots, n\}$, for $L \geq 1$, a candidate set of CCPs satisfying $\mathcal{T} \subseteq \mathcal{K}$. Moreover, denote for all $\ell \in \{1, \dots, L\}$, the intervals $I_{k_\ell} \coloneqq \{k_{\ell-1}, \dots, k_{\ell+1} - 1\}$, where $k_0 = 1$ and $k_{L+1} = n+1$. Let $\widehat{\mathcal{T}}$ be the CCP estimator defined in Algorithm 2 and let $(\phi_I)_{I \in \mathcal{I}}$ be a collection of tests for $(\mathcal{H}_0^I)_{I \in \mathcal{I}}$. Then, the following two statements hold:*

*(i) Let $\alpha \in (0,1)$. If for all $k \in \mathcal{K}$ it holds that $\phi_{I_k}$ is level $\alpha$, then*

$$\mathbb{P}(\widehat{\mathcal{T}} \subseteq \mathcal{T}) \geq 1 - (|\mathcal{K}| - |\mathcal{T}|) \cdot \alpha.$$

*(ii) Let $\beta \in (0,1)$. If for all $\ell \in \{1, \dots, L\}$ with $k_\ell \in \mathcal{T}$ it holds that $\mathbb{P}(\phi_{I_{k_\ell}} = 1) \geq \beta$, then*

$$\mathbb{P}(\mathcal{T} \subseteq \widehat{\mathcal{T}}) \geq 1 - |\mathcal{T}| \cdot (1 - \beta).$$

A proof can be found in Appendix 3.D. Following Proposition 3.4.1, one may adjust $\alpha$ by a factor $c \leq 1/(|\mathcal{K}| - |\mathcal{T}|)$ which ensures that $\mathbb{P}(\widehat{\mathcal{T}} \subseteq \mathcal{T}) \geq 1 - \alpha$. One special case is the Bonferroni correction, which corresponds to $c = 1/|\mathcal{K}|$ and always preserves coverage at level $\alpha$ but might be conservative if there are many CCPs. In practice, the candidate set may not be a superset of the true CCPs, i.e., the candidate set may not contain all the true CCPs, or some candidates are time points that slightly deviate from the true CCPs, or a combination of both. If this is the case, the resulting CCP estimates can be arbitrarily biased. To check the validity of the estimates, one can test each of

the sub-intervals separated by $\widehat{\mathcal{T}}$: if there is no invariant set in a sub-interval, it means that there exist at least one CCP that is not in the candidate set, or the candidates surrounding the sub-interval are not true CCPs. We illustrate this in Appendix 3.C.1.

### 3.4.2 Causal change point localization via a loss function

An alternative approach to localizing causal change points is by finding the minima of a loss function. Here, we propose a loss function that assesses the level of causal non-invariance at each time point. Ideally, the loss function should achieve its minimal value in an interval with a single CCP at the true CCP. In Section 3.4.2.1, we introduce the loss function and discuss its properties at population level given a single CCP. We discuss how to estimate the location of one CCP using an empirical version of the loss function in Section 3.4.2.2. Localization of multiple CCPs is discussed in Section 3.4.2.3, where we leverage modified versions of existing multiple change point detection algorithms [Vostrikova, 1981, Baranowski et al., 2019, Kovács et al., 2023] to localize each of them.

#### 3.4.2.1 Causal stability loss at population level

In this section, we introduce a loss function to capture the change in causal mechanism of the response $Y$. Intuitively, for an interval $I \in \mathcal{I}$, the loss at a time point $i \in I$ sums up the level of non-invariance over the two sub-intervals of $I$ to the left and right of $i$. Suppose there exists exactly one CCP in $I$, this loss function achieves its minimum value at the true CCP at population level. To formally introduce the loss function, we require the following notation.

**Notation 3.4.1.** *Let $s \in \mathbb{N}$ be a minimal segmentation length. For all intervals $I \in \mathcal{I}$ with $|I| \geq 2s$, let $m_s(I) := \lfloor \frac{|I|}{s} \rfloor$ and let $P_1(I), \dots, P_{m_s(I)}(I)$ be a partition of $I$ into $m_s(I)$ intervals such that $|P_r(I)| = s$ for $r \in \{1, \dots, m_s(I) - 1\}$ and $|P_{m_s(I)}(I)| = |I| - \big(m_s(I) - 1\big) \cdot s$. For all $r \in \{1, \dots, m_s(I)\}$, we denote the complement of $P_r(I)$ as $P_r^c(I) := I \setminus P_r(I)$. For all $I \in \mathcal{I}$ with $|I| < 2s$, we let $m_s(I) = 1$, $P_1(I) = I$, and with a slight abuse of notation, $P_1^c(I) = I$. An illustration of this notation is given in Figure 3.4.3. For all $I, J \in \mathcal{I}$, we define $V_{I,J}(S) = \frac{1}{|I|} \sum_{\ell \in I} \mathbb{E}\Big[\big(\epsilon_\ell^J(S)\big)^2\Big]$ where $\epsilon_\ell^J(S)$ is the population OLS residual at $\ell$ given $S$ over $J$ (see Definition 3.3.2).*

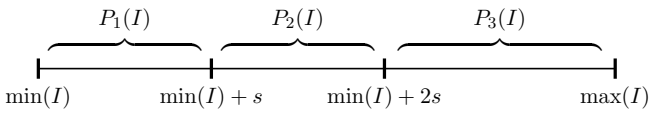



Figure 3.4.3: Illustration of the partition of an interval $I$ with $3s \leq |I| \leq 4s - 1$ into 3 sub-intervals. The ticks $\min(I)$, $\min(I) + s$ and $\min(I) + 2s$ mark the beginning of the three intervals and $\max(I)$ marks the end of the third interval.

If there is no CCP in $I \in \mathcal{I}$, then there exists $S \in \mathcal{S}$ and $c \in \mathbb{R}$ such that for all $J \subseteq I$, $V_{I\setminus J,J}(S) = V_{J,J}(S) = c$. This motivates the following definition of minimal OLS instability (Definition 3.4.2) which serves as the basis of our causal stability loss (Definition 3.4.4).

**Definition 3.4.2** (Minimal OLS instability)**.** Let $I \in \mathcal{I}$ and let $s \in \mathbb{N}$ be a minimal segmentation length. The *minimal OLS instability* over the interval $I$ is defined as

$$\mathcal{C}_s(I) \coloneqq \min_{S\in\mathcal{S}} \sum_{r=1}^{m_s(I)} \left(V_{P_r^c(I),P_r(I)}(S) - V_{P_r(I),P_r(I)}(S)\right)^2.$$

♣

It satisfies the following property.

**Proposition 3.4.3.** *Let $I \in \mathcal{I}$ and $s \in \mathbb{N}$. Suppose there is no CCP in $I$, then $\mathcal{C}_s(I) = 0$.*

A proof is given in Appendix 3.D. We then define the causal stability loss at a time point $i$ in an interval $I$ as the sum of minimal OLS instability over the sub-intervals to the left and right of the time point $i$.

**Definition 3.4.4** (Causal stability loss)**.** Let $I \in \mathcal{I}$ and let $s \in \mathbb{N}$ be a minimal segmentation length. For all $i \in I \setminus \{\min(I), \max(I)\}$, we define $I_{i-} \coloneqq \{\min(I), \dots, i-1\}$ and $I_{i+} \coloneqq \{i, \dots, \max(I)\}$, then we define the *causal stability* loss as

$$\mathcal{L}_{I,s}(i) = \frac{\mathcal{C}_s(I_{i-}) + \mathcal{C}_s(I_{i+})}{m_s(I_{i-}) + m_s(I_{i+})},$$

where $\mathcal{C}_s(I_{i-})$ and $\mathcal{C}_s(I_{i+})$ are as defined in Definition 3.4.2. ♣

The following property of the causal stability loss follows directly from Proposition 3.4.3.

**Corollary 3.4.5.** *Let $I \in \mathcal{I}$ and $s \in \mathbb{N}$. If there is no CCP in $I$, then for all $i \in I$ $\mathcal{L}_{I,s}(i) = 0$; if $\tau \in \{\min(I)+1, \dots, \max(I)-1\}$ is the only CCP in $I$, then $\mathcal{L}_{I,s}(\tau) = 0$.*

#### 3.4.2.2 Localizing a single causal change point

The loss function $\mathcal{L}_{I,s}$ can be estimated by replacing the population quantities with their empirical counterparts. The OLS coefficient given $S \in \mathcal{S}$ over an interval $I \in \mathcal{I}$ can be estimated by the vector $\hat{\beta}_I^{\mathrm{OLS}}(S) \in \mathbb{R}^{d+1}$ where

$$\left(\hat{\beta}_I^{\mathrm{OLS}}(S)\right)^S = \operatorname*{arg\,min}_{\beta^S\in\mathbb{R}^{|S|}} \sum_{\ell\in I} \left(Y_\ell - (X_\ell^S)^\top \beta^S\right)^2$$

and $\left(\hat{\beta}_I^{\mathrm{OLS}}(S)\right)^j = 0$ for all $j \in \{1, \dots, d+1\} \setminus S$. For all $I \in \mathcal{I}$, let $\hat{\beta}_I^{\mathrm{OLS}}(S)$ be the estimated OLS coefficient given $S$ over $I$. For all $\ell \in \{1, \dots, n\}$, the estimated OLS

residual at $\ell$ given $S$ over $I$ is given by $\hat{\epsilon}_\ell(S) = Y_\ell - X_\ell^\top \hat{\beta}_I^{\mathrm{OLS}}(S)$ with the convention that $\hat{\epsilon}^I(S) = \left(Y_\ell - X_\ell^\top \hat{\beta}_I^{\mathrm{OLS}}(S)\right)_{\ell \in I} \in \mathbb{R}^{|I|}$. Lastly, for all $I, J \in \mathcal{I}$, let $\widehat{V}^2_{I,J}(S) := \frac{1}{|I|}\sum_{\ell \in I}\left(\hat{\epsilon}_\ell^J(S)\right)^2$. Then the minimal OLS instability over $I$ can be estimated by

$$\widehat{\mathcal{C}}_s(I) := \min_{S \in \mathcal{S}} \sum_{r=1}^{m(I)} \left(\widehat{V}_{P_r^s(I),P_r(I)}(S) - \widehat{V}_{P_r(I),P_r(I)}(S)\right)^2$$

and the causal stability loss can be estimated by

$$\widehat{\mathcal{L}}_{I,s}(i) := \frac{\widehat{\mathcal{C}}_s(I_{i-}) + \widehat{\mathcal{C}}_s(I_{i+})}{m(I_{i-}) + m(I_{i+})}.$$

#### 3.4.2.3 Localizing multiple causal change points

We propose two general approaches to localize multiple CCPs. The first approach uses the standard binary segmentation algorithm proposed by Vostrikova [1981] (see Algorithm 3 in Appendix 3.B) and then prunes the resulting estimates by Algorithm 2 in Appendix 3.B. The pruning step is necessary since even at population level, the causal stability loss does not necessarily achieve its minimum at one of the true CCPs in an interval that contains multiple CCPs, although it does so when only one CCP exists in an interval.

As a second approach, we consider a different greedy algorithm which instead of searching for change points in a top-down order as the standard binary segmentation, it searches for change points in a bottom-up order, namely the seeded binary segmentation algorithm [Kovács et al., 2023] with the narrowest-over-threshold selection procedure [Baranowski et al., 2019]. The idea is to first generate a collection of sets of intervals with increasing lengths[2]. Among the narrowest intervals for which $\mathcal{H}_0^I$ is rejected, we estimate one CCP in the interval that has the smallest p-value, and eliminate all intervals that contain the estimated CCP. We then repeat the procedure among the remaining sets of intervals from the narrowest to the widest until $\mathcal{H}_0^I$ is not rejected for any remaining intervals. The bottom-up order aims to ensure that each interval only contains at most one CCP, which is suitable when the loss function may not achieve its minimum at a CCP given multiple CCPs in an interval. The procedure of obtaining the seeded intervals [Definition 1, Kovács et al., 2023] is given in Algorithm 4 in Appendix 3.B. Algorithm 5 in Appendix 3.B describes the overall procedure of the second approach. We compare the above approaches in Section 3.5.1.

## 3.5 Numerical Experiments

We demonstrate the performance of our proposed methods based on both simulated datasets and two real datasets. In Section 3.5.1 we describe the data generating process

[2]The seeded intervals can be seen as generated layer by layer, as described in Algorithm 4. The intervals on the same layer are considered to have the same length, despite the small differences caused by rounding.

that the simulated experiments are based on. For CCP detection, we show the level and power given differet true locations of one CCP with the Chow test for testing $\mathcal{H}^I_{0,S}$, as described in Section 3.3.1. For CCP localization, we consider both the case where it is known that there is exactly one CCP and the case where there are multiple CCPs, and compare the methods proposed in Section 3.4. Additional numerical experiments can be found in Appendix 3.C.1. All numerical experiments can be reproduced using the code available at `https://github.com/shimenghuang/CausalCP`.

### 3.5.1 Simulated experiments

In this section, we consider the following data generating process given for all $i \in \{1, \dots, n\}$ by

$$\begin{aligned} X_i^1 &:= \epsilon_i^1 \\ X_i^2 &:= \alpha_i^{12} X_i^1 + \epsilon_i^2 \\ Y_i &:= \beta_i^{15} X_i^1 + \beta_i^{25} X_i^2 + \epsilon_i^Y \\ X_i^4 &:= \epsilon_i^4 \\ X_i^3 &:= \alpha_i^{53} Y_i + \alpha_i^{43} X_i^4 + \epsilon_i^3, \end{aligned} \tag{3.5.8}$$

where $\epsilon_i^j \sim \mathcal{N}(\mu_i^j, (\sigma_i^j)^2)$ for $j \in \{1, 2, 3, 4, Y\}$. The induced DAG for all $i \in \{1, \dots, n\}$ is shown in Figure 3.5.4. There are in total 15 parameters in this data generating process which can be divided into two sets: 4 parameters that are related to the causal mechanism of $Y$ with respect to $X$ ($\beta_i^{15}$, $\beta_i^{25}$, $\mu_i^Y$, and $\sigma_i^Y$), and 11 parameters that are not related to causal mechanism of $Y$ with respect to $X$ ($\alpha_i^{12}$, $\alpha_i^{53}$, $\alpha_i^{43}$, $\mu_i^j$ for $j \in \{1, 2, 3, 4\}$, and $\sigma_i^j$ for $j \in \{1, 2, 3, 4\}$). Throughout this section, we refer to the set of parameters that are related to the causal mechanism of $Y$ as the *causal parameters*, and the set of parameters thar are not related to causal mechanism of $Y$ as *non-causal parameters*. In the following experiments, parameters are chosen such that changes in the causal parameters are CCPs (see Definition 3.2.4) and changes in the non-causal parameters are NCCPs (we have verified this using straight-forward computations). More details of the data generating process can be found in the Appendix 3.C.2. All experiments are based on 200 repetitions.

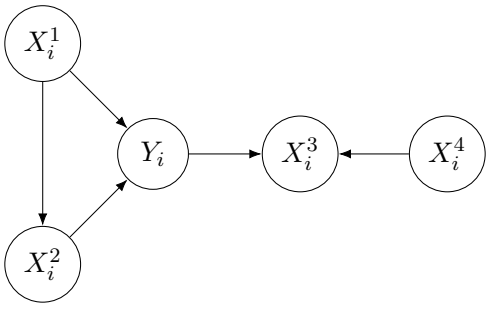



Figure 3.5.4: DAG induced by (3.5.8) for all $i \in \{1, \dots, n\}$.

#### 3.5.1.1 CCP detection

We demonstrate the power to detect CCPs where changes happen in the causal parameters $\beta_i^{15}$ and $\beta_i^{25}$ in (3.5.8) of the split Chow test described in Section 3.3.1 and show that it holds the correct level. We fix $\alpha$ to be 0.05 in this experiment. Figure 3.5.5 shows that the procedure has the most power when the true CCP is in the middle of the interval while it has least power when the true relative location of one single CCP is close to the boundaries of the interval. A possible explanation is that the Chow test is applied on the two sub-intervals to the left and right of the midpoint. When the true CCP is to the left of the midpoint, the left half of the interval contains data from a mixture of two distributions before and after the change, and the right half contains data only from the distribution after the change, similarly when the true CCP is to the right of the midpoint. Only when the true CCP coincides with the midpoint, the two sides both contain data from a single distribution which leads to a higher power with the Chow test. The "no CCP" label on the x-axis corresponds to when the interval does not contain a CCP. In that case a valid method should control the type-I error of wrongly detecting a CCP at the pre-specified level (5% in this case).

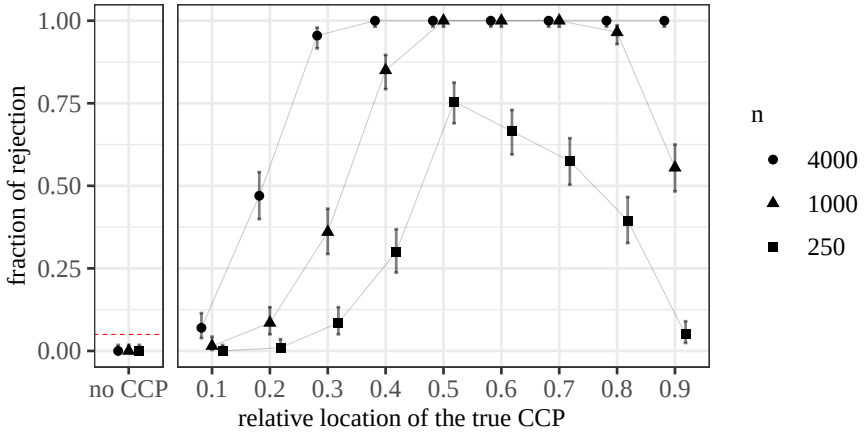



Figure 3.5.5: Empirical investigation of level and power of the testing procedure described in Section 3.3.1 with increasing sample sizes. The x-axis corresponds to the relative location of a single CCP (and no CCP). The error bars are binomial confidence intervals and the red dashed line is at 0.05.

#### 3.5.1.2 CCP localization

We compare the proposed methods for localizing CCPs described in Section 3.4 and the method `BP` which uses the 'breakpoints' method for localizing structural changes in linear models due to Bai and Perron [2003] implemented in the R package 'strucchange' [Zeileis et al., 2002]. In Experiment 1 and Experiment 2, which consider datasets with a single true CCP, we apply the pruning approach described in Section 3.4.1 assuming the set of true RCPs is known (`Prune-Oracle`) and assuming that they are unknown then

using breakpoints to estimate the RCPs (`Prune-BP`). These approaches are compared with the causal stability loss (`LossCS`) approach described in Section 3.4.2 and using breakpoints to estimate a single change point (`BP-1`). In Experiment 3 where there are two true CCPs, we apply the following methods: (i) using breakpoints to estimate two change points (`BP-2`), (ii) using breakpoints to estimate all change points up to a specified minimal segment length (`BP`), (iii) using seeded binary segmentation with the causal stability loss as in Algorithm 5 in Appendix 3.B (`LossCS-SeededBS`), (iv) using `LossCS-SeededBS` combined with a pruning step (`LossCS-SeededBS-Prune`), (v) using standard binary segmentation with the causal stability loss as in Algorithm 3 in Appendix 3.B (`LossCS-StdBS`), and (vi) using `LossCS-StdBS` combined with a pruning step (`LossCS-StdBS-Prune`).

**Experiment 1: Fixed relative locations of one CCP and two NCCPs with increasing $n$** For a total number of time points $n$, one true CCP is fixed at $0.5n + 1$ and two true NCCPs are fixed at $\lceil 0.25n+1 \rceil$ and $\lceil 0.75n+1 \rceil$, respectively. The minimal segmentation length for all methods is set to be $0.1n$. For `LossCS`, we evaluate the loss at every $0.05n$ time points starting from $\lceil 0.05n + 1 \rceil$ and ending at $\lceil 0.95n + 1 \rceil$. We compare the different methods for sample sizes $n \in \{250, 1000, 4000\}$. In Figure 3.5.6, we see that under this setup, all methods except `BP-1` give estimates concentrating around the true value with increasing sample size, while estimates based on `BP-1` concentrate around the second NCCP. Moreover, `LossCS` performs better than the two pruning approaches `Prune-Oracle` and `Prune-BP` which both end up not detecting any CCPs in many simulations. Lastly, the results of `Prune-Oracle` and `Prune-BP` are similar for large sample sizes, indicating that estimating RCPs from data and using these estimates as candidates can indeed perform well.

**Experiment 2: Different relative locations of a single CCP** For $n = 2000$, two true NCCPs are fixed at $0.25n + 1$ and $0.75n + 1$, respectively. We evaluate the performance of `LossCS` when the relative location of a single CCP is fixed at $\nu n + 1$ for $\nu \in \{0.1, 0.2, \ldots, 0.9\}$. We fix the minimum segmentation length to be $0.1n$ and for `LossCS` we evaluate the loss at every $0.1n$ points. Figure 3.5.7 shows the estimated versus true relative locations of the CCP. The points are jittered horizontally for visual clarity. We can see that `LossCS` performs well when the true CCP is relatively close to the center of the interval, but when the true CCP is close to the left (respectively, right) boundary of the interval, `LossCS` tends to over- (respectively, under-) estimate the location.

**Experiment 3: Fixed relative locations of two CCPs and one NCCP with increasing $n$** For a total number of time points $n$, two CCPs are fixed at $0.2n+1$ and $0.8n+1$, and one NCCP is fixed at $0.5n + 1$. We fix the minimum segmentation length for all methods to be $0.2n$ and for `LossCS` we evaluate the loss at every time point other than the $\max(\lceil 0.1|I| \rceil, 10)$ points at the beginning and end of each seeded interval $I$. We compare the different methods for sample sizes $n \in \{1000, 2000, 4000\}$. Bonferroni correction is

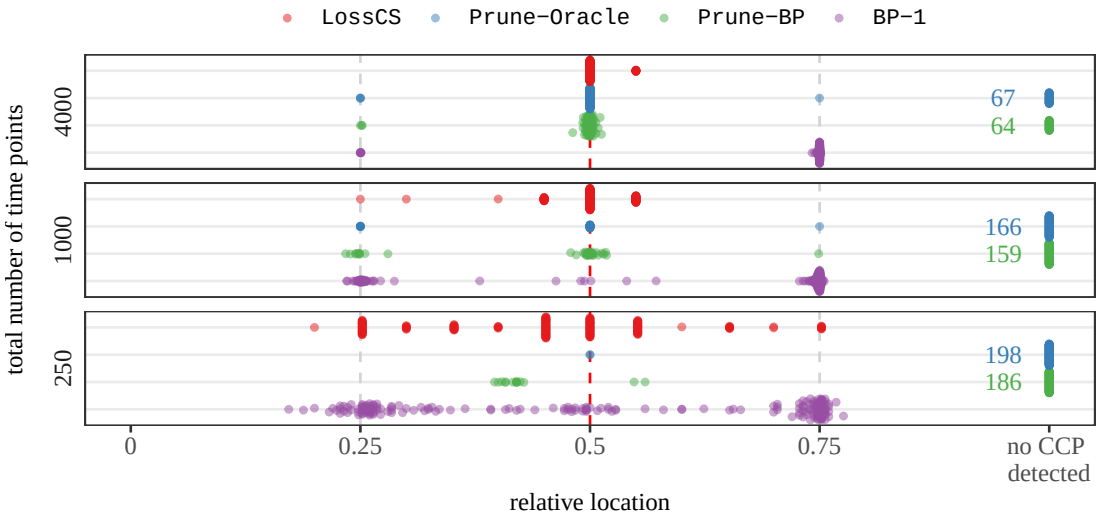



Figure 3.5.6: Relative locations of the estimated CCP using different methods with varying number of time points $n$. The red vertical dashed line indicates the true relative location of the CCP and the two grey vertical dashed lines correspond to the true relative locations of the NCCPs. With increasing sample size, estimates from all methods except `BP-1` concentrate around the true CCP, but `Prune-Oracle` and `Prune-BP` both detect no CCP in many of the 200 repetitions.

applied in the pruning step when there is more than one candidate. Figure 3.5.8 contains the histogram of the estimates based on each approach under this setting. As can be seen, using `BP-2` to estimate two CCPs is not a valid method as it might detect NCCPs as in this example. With a large enough sample size, `Prune-BP` performs best based on the number of false and true positives, followed by `LossCS-SeedBS-Prune`, `LossCS-SeedBS`, and `LossCS-StdBS-Prune`. When the sample size is small, `LossCS-SeedBS-Prune` performs best in the sense that it has the least number of false positives and the most number of true positives. The pruning step after `LossCS-SeedBS` does not seem to improve the estimation much especially when the sample size is large, as the number of false positives is already low. However, for `LossCS-StdBS` the pruning step is crucial in this example, as it tends to first split on the NCCP hence leading to many false positives.

In summary, both families of `LossCS-*` and `Prune-*` can be helpful when localizing CCPs. The `LossCS-*` methods can be used in combination with many existing change point localization schemes. The `Prune-*` methods come with the usual guarantees of a test, which may be beneficial for small sample sizes (where it may be better to remain conservative and not make any decision). If the set of candidates is incorrect in that it does not contain all true CCPs, the output of these methods may be incorrect; however, in some scenarios it may be possible to realize that, see Experiment 4 in Appendix 3.C.1.

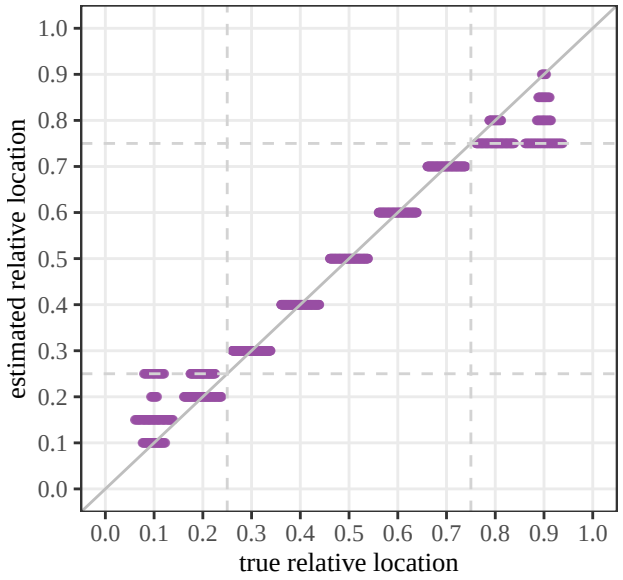



Figure 3.5.7: Estimated relative locations the CCP given different true relative locations of the CCP and $n = 2000$. The two vertical dashed lines indicate the (fixed) true relative locations of the NCCPs. `LossCS` tends to over- (respectively, under-) estimate the location of CCP when the CCP is close to the left (respectively, right) boundary.

### 3.5.2 Real data application

We now consider two real datasets and use them to illustrate the differences between CCPs and other types of change points. The first dataset, which we use to illustrate CCP detection and localization, pertains to air quality in Beijing during February 2014 and March 2017. It is taken from the UCI Machine Learning Repository [Chen, 2017] and was previously analyzed by Zhang et al. [2017]. The second dataset, which we use to illustrate CCP detection, is the monetary policy example discussed by Pfister et al. [2019]. The purpose of these examples is only to illustrate possible scenarios that our methods can be applied to, rather than to provide new subject-matter insights based on the two datasets.

To account for time dependence and detect and localize changes in the instantaneous causal effects, we adopt the approach proposed in Pfister et al. [2019], which relies on linear autoregressive models. To be more precise, this means that (3.2.2b) in Setting 1 is replaced by the following:

$$Y_i = X_i^\top \beta_i + \sum_{l=1}^{\ell} (Y_{i-l}, X_{i-l}^\top) b_l + g_i(H_i, \epsilon_i^Y),$$

where $\ell \in \{1, \dots, n-2\}$ is a fixed number of lags[3], and $b_l \in \mathbb{R}^{d+1}$ for $l \in \{1, \dots, \ell\}$.

[3]We consider $\ell$ to be at least 1 because 0 lag reduces to the case without any time dependence.

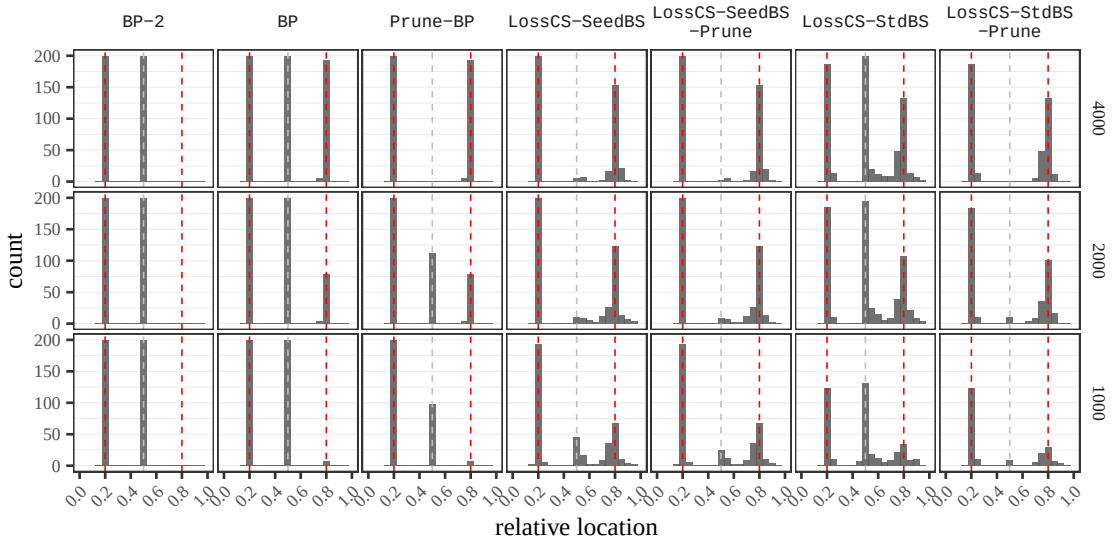



Figure 3.5.8: Histogram of the estimated CCPs using different methods. The two red dashed lines correspond to the relative location of the true CCPs. The grey dashed line corresponds to the relative location of the true NCCP. When the sample size is large, `Prune-BP` performs best in terms of both the number of false positives and true positives while at a relatively small sample size `LossCS-SeedBS-Prune` performs best.

This modified setting is satisfied, for example, if $\{(X_i, Y_i)\}_{i=1}^n$ follows a structural vector autoregressive process [see e.g., Lütkepohl, 2005].

In terms of methodology, this requires only an augmentation of the covariates $X$. Given a set of observations $\{(Y_i, X_i)\}_{i=1}^n$, our CCP detection and localization methods are consequently adapted as follows: for a set $S \in \mathcal{S}$ and an interval $I \in \mathcal{I}$, instead of regressing $Y_i$ on $X_i^S$, we regress $Y_i$ on the augmented predictors $Z_i^S := \big((X_i^S)^\top, Y_{i-1}, X_{i-1}^\top, \ldots, Y_{i-l}, X_{i-l}^\top\big)^\top$ for a fixed number of lags $\ell \in \{1, \ldots, |I| - 2\}$. In the following applications, we report the results using different numbers of lags.

### 3.5.2.1 Air quality example

The air quality dataset comprises hourly observations of variables related to air pollutants and meteorological conditions in Beijing. Our analysis focuses on a monitoring site located in the city center of Beijing. We aggregate the hourly data for PM2.5, SO2, NO2, and CO concentrations, as well as meteorological variables DEWP and WSPM, into daily averages over the period from January 1, 2014, to January 1, 2017. A brief description of these variables is provided in Appendix 3.C.3.

PM2.5 is an air pollutant that poses a serious health risk to the population, as particles and droplets of this size can penetrate deep into the lungs and even enter the bloodstream [see e.g., Thangavel et al., 2022]. Since 2013, Beijing has established an air pollution

monitoring network as part of its efforts to address air quality issues [Cheng et al., 2019]. The analysis by Zhang et al. [2017] shows that although the average PM2.5 concentration decreased in 2015 both marginally and after adjusting for meteorological factors, in 2016, it only decreased marginally but not after adjusting for meteorological factors. On the other hand, the analysis by Cheng et al. [2019] shows that although the meteorological factors influence the PM2.5 level to a large extent, the emission reduction measures during 2013 to 2017 do contribute to controlling the PM2.5 level. We apply our CCP detection method to investigate whether there is a change in how PM2.5 causally depends on the covariates—which may indicate a significant intervention in the PM2.5 emission mechanism—and compare it with marginal change point (MCP) and RCP detection results.

We log-transform the four pollutant concentration variables. The log PM2.5 concentration is considered as the response, and the log concentration of the other three pollutants, two meteorological variables, and one heating indicator are taken as the covariates. We take heating into account as coal and gas are used to provide winter heating in Beijing, which likely contributes to the seasonal pattern in the concentration of air pollutants. The heating indicator is set to 1 during November 15 to March 15 each year according to Beijing's heating schedule [Beijing Municipal Government, 2009]. As time dependence is expected, we include 1 to 7 lags of all covariates and the response in the linear autoregressive model. MCP detection is conducted by the CUSUM test [Page, 1955] and RCP detection is done using the Chow test [Chow, 1960].

The marginal distribution of log-transformed PM2.5 is shown in the left plot of Figure 3.5.9. We can see the distinct patterns of PM2.5 concentration during the winter heating season compared to other times of the year. This is consistent with MCP detection results in the right plot of Figure 3.5.9. For RCP and CCP, we first consider including all six covariates in the model. From the top plot in Figure 3.5.9, we see that the hypotheses of no MCP and no RCP are rejected at a 5% level for all 1 to 7 lags, while no CCP is not rejected for any lags, showing that the data provides evidence for the existence of MCPs and RCPs but not for CCPs. We do detect all three types of change points, however, if we only include the two meteorological variables as covariates, as shown in the bottom plot in Figure 3.5.9. This indicates that the omitted variables possibly have a causal effect on the concentration of PM2.5, and their distributions might have changed during the time interval. We apply our CCP localization method (`LossCS-SeedBS`) to localize the CCPs with the meteorological-variable-only model, and compare with RCP and MCP localization results using the 'breakpoints' method (BP) in the 'strucchange' R package [Zeileis et al., 2002, Bai and Perron, 2003], and wild binary segmentation (WBS) by Fryzlewicz [2014], respectively. The localized change points (dates) are given in Table 3.5.1. We can see that although the CCP estimates are not exactly the same when including different numbers of lags, the estimated dates are consistently around the start or end of the winter heating period. This suggests that winter heating, which is not accounted for in the meteorological-variable-only model and alternates on and off each year, appears (as one would expect) to lead to causal changes in the PM2.5 concentration. In contrast, RCP and MCP localization methods result in many more change points throughout the years.

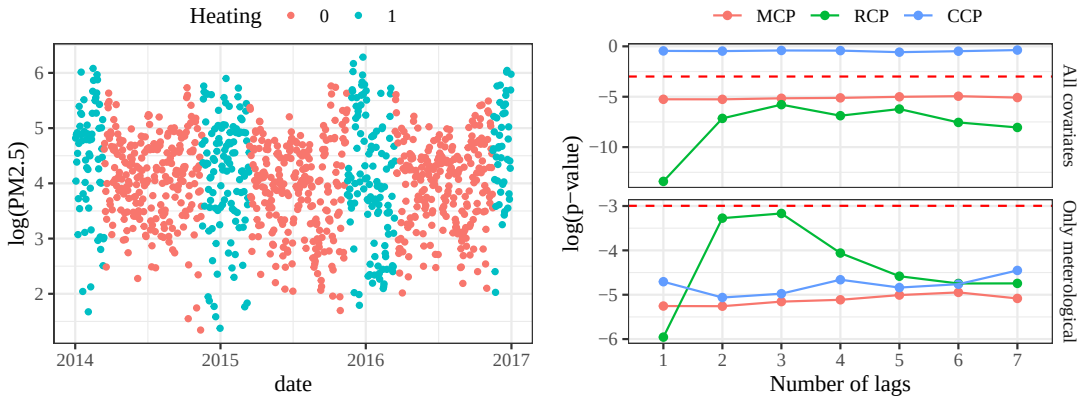



Figure 3.5.9: Left: Marginal distribution of log PM2.5 concentration, colored by whether the date is during winter heating or not. Right: p-values obtained from MCP, RCP, and CCP detection tests. 1 to 7 lags of the covariates and response are included in the regression model. The red dashed lines indicate the rejection threshold at log(0.05).

| Number of lags | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2014-11-05 | 2014-11-02 | 2014-11-03 | 2014-11-11 | 2015-04-13 | 2014-11-11 | 2014-11-10 |
| 2015-03-05 | 2016-04-22 | 2015-03-18 | 2015-10-24 | 2015-10-20 | 2015-04-28 | 2015-04-27 |
| 2016-04-26 | | 2015-11-04 | 2016-04-21 | 2016-04-27 | 2015-10-23 | 2016-04-25 |
| | | 2016-04-22 | | | 2016-04-26 | |

Table 3.5.1: Localized CCPs using `LossCS-SeedBS` based on the meteorological-variable-only model when including 1 to 7 lags. All CCPs are localized around the beginning and end of winter heating period. In contrast, we detect 43 MCPs based on WBS (for MCPs, we do not regress on the past, so no lags are involved), and for all lags, we detect 8 RCPs. We report the dates of the estimated MCPs and RCPs in Appendix 3.C.3.

### 3.5.2.2 Monetary policy example

The monetary policy dataset consists of monthly observations of 10 variables related to the monetary policy of the Swiss National Bank (SNB) from January 1999 to January 2017. As response $Y$, we consider the Swiss franc (CHF)/Euro (EUR) exchange rate, which is driven by economic factors and the monetary policy of the SNB both of which are at least partially captured by the remaining 9 variables $X^1, \ldots, X^9$. A short description of all variables and their preprocessing is provided in Appendix 3.C.3. In the observed time frame there were two major events that affected the exchange rate: (i) The SNB enforced a cap on the value of the Swiss franc of 1.2 CHF/EUR between September 2011 and January 2015 by trading the necessary amount of foreign currency. This change in

the policy is captured in variable $X^2$, the log returns of end of month proportion of foreign currency investments from total assets on the balance sheet of the SNB. (ii) The 2009 Greek debt crisis negatively affected the value of the Euro and increased the value of the Swiss franc. The inflow of money into Switzerland during this crisis is captured (at least to some degree) by the Swiss gross domestic product measured by $X^7$. To fully model both events, the additional variables both on the economy ($X^8$) and the policy of the SNB (all remaining variables) are likely to be useful.

We now apply our CCP detection method to determine whether there is a CCP in how the exchange rate depends on the explanatory variables. The resulting p-values for a CCP and RCP detection test with $\ell \in \{2, 3, 4\}$ lags are shown in Table 3.5.2 (under 'All covariates'). From the discussion above, we do not expect a CCP to be present because the model contains all factors relevant to the response. The CCP detection algorithm (using the 'variance' invariance test proposed in Pfister et al. [2019]) indeed does not reject the null hypothesis of no CCP at a 5% level. Furthermore, if we instead test for the existence of an RCP using the same procedure but only testing $H_0^{\{1,\dots,d+1\}}$, we reject the null hypothesis at a 5% level, indicating the existence of an RCP.

To further illustrate the difference between CCPs and RCPs, we now consider the same data but remove $X^2$, $X^7$, or both $X^2$ and $X^7$ from the analysis. As discussed above, we believe that these are relevant factors for explaining the Greek debt crisis and the exchange rate cap. The results are again shown in Table 3.5.2 (under 'Without $X^2$', 'Without $X^7$', and 'Without $X^2$ and $X^7$'). As expected, the CCP detection test now rejects the null hypothesis of no CCPs in all three cases, indicating that there was a shift that cannot be explained by the covariates used in the model.

| p-values | All covariates | | | Without $X^2$ | | | Without $X^7$ | | | Without $X^2$ and $X^7$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lags | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 |
| no RCP | 0.015 | 0.015 | 0.020 | 0.006 | 0.002 | 0.005 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| no CCP | 0.045 | 0.104 | 0.090 | 0.014 | 0.014 | 0.019 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

Table 3.5.2: p-values for testing the null hypotheses that there is no RCP and there is no CCP using all covariates and using all covariates except $X^2$, $X^7$, or $X^2$ and $X^7$. With all covariates and given sufficiently many lags we only detect RCPs. When omitting variables $X^2$ and $X^7$ we detect both RCPs and CCPs at level $\alpha = 0.05$.

## 3.6 Discussion

We introduce a notion of causal changes which, under additional causal assumptions, captures changes in the underlying causal mechanism, but which is still meaningful without referencing an underlying causal model. We consider the problems of detecting and localizing these changes in an offline sequential setting. For detection, we propose a testing procedure that combines several invariance tests. For localization, we propose

two approaches based on pruning a set of candidate CCPs and based on minimizing a loss function, respectively. The first approach `Prune-*` is directly applicable if a superset of all CCPs is known. If this is not the case, one can first estimate the RCPs using a method that can localize both changes in the regression parameter and changes in the residual distribution. If the candidates are imprecise, an NCCP may be mistaken as a CCP in the final estimates. The second approach `LossCS-*` avoids this problem by targeting the CCPs directly without estimating the RCPs first.

In the Beijing air quality and monetary policy examples, we demonstrate the possibility to apply our CCP detection method to detect instantaneous causal changes with time dependencies. One potential future direction is to consider also the causal mechanism relating the response and the covariates' past such as detecting and localizing changes in the number of lags of the covariates. Moreover, we have focused on linear regression models, but the ideas can potentially be generalized to semi-parametric or non-parametric regression models. In those settings, we may rely on a notion of invariant functions [e.g., Christiansen et al., 2021] rather than invariant sets. Lastly, it may also be of interest to quantify uncertainty for the CCP localization problem, see for example, Fryzlewicz [2024].

## Acknowledgement

SH and NP are supported by a research grant (0069071) from Novo Nordisk Fonden. We thank Solt Kovács for the helpful discussions and Rikke Søndergaard for her contribution in the initial stage of this work. Part of the work was done while JP was at the University of Copenhagen.

# Supplementary material for "Causal change point detection and localization"

- Section 3.A: Additional examples and details on examples.
- Section 3.B: Algorithms
- Section 3.C: Additional numerical experiments and experiment details
- Section 3.D: Proofs
- Section 3.E: Auxiliary results
- Section 3.F: Chow test

## 3.A. Additional examples and details on examples

### 3.A.1 Details of Example 3.2.2

For all $j \in \{X^1, X^2, X^3, Y\}$, denote the variance of $\epsilon_i^j$ as $(\sigma_i^j)^2$. For all $i \in \{1, \dots, n\}$, $\alpha = (0, 0, 1)^\top$; for all $i \in I_1 \cup I_2$, $A_i = \mathbf{0}_{3\times 4}$ and $\beta_i = (1, 1, 0, 0)^\top$, and for all $i \in I_3$, $A_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$ and $\beta_i = (0, 1, 0, 0)^\top$. For all $i \in I_1$ and for all $j \in \{X^1, X^2, X^3, Y\}$, $\sigma_i^j = 1$, and for all $i \in I_2 \cup I_3$, $\sigma_i^{X^1} = \sigma_i^Y = 1$ and $\sigma_i^{X^2} = \sigma_i^{X^3} = 2$. This means that the population OLS coefficient for all $i \in I_1$ is $\beta^{\mathrm{OLS}} = (0.5, 0.5, 0.5, 0)^\top$, for all $i \in I_2$ is $\beta^{\mathrm{OLS}} = (0.8, 0.8, 0.2, 0)^\top$, and for all $i \in I_3$ is $\beta^{\mathrm{OLS}} = (-0.2, 0.8, 0.2, 0)^\top$.

### 3.A.2 Example where the reverse implication of Proposition 3.2.7 does not hold

For all $i \in \{1, \dots, n\}$ consider an SCM over $(X_i, Y_i)$ given by $X_i^3 := 1$ as intercept and

$$\begin{aligned} X_i^1 &:= \epsilon_i^{X^1}, \\ Y_i &:= \beta_i X_i^1 + \epsilon_i^Y, \\ X_i^2 &:= \alpha_i Y_i + \epsilon_i^{X^2}, \end{aligned}$$

where $\epsilon_i^j \sim \mathcal{N}\left(0, (\sigma_i^j)^2\right)$ for all $j \in \{X^1, X^2, Y\}$. Let $k \in \{1, \dots, n\}$ with $1 < k < n$. For all $i \in \{1, \dots, k-1\}$, $\beta_i = 2$, $\alpha_i = 3$, and for all $j \in \{X^1, X^2, Y\}$, $\sigma_i^j = 1$. For all $i \in \{k, \dots, n\}$, $\beta_i = 1$, $\alpha_i = 8/3$, $\sigma_i^{X^1} = 1$, $\sigma_i^{X^2} = \sqrt{2}$, and $\sigma_i^Y = 3\sqrt{2}/4$ . Then, for all $i \in \{1, \dots, n\}$, $\hat{\beta}_i^{\mathrm{OLS}} = (0.2, 0.3, 0)^\top$. Thus, $k$ is not a CCP even though $\beta_k \neq \beta_{k-1}$ and $\epsilon_k^Y \overset{d}{\neq} \epsilon_{k-1}^Y$.

## 3.B. Algorithms

The procedure of testing candidates described in Section 3.4.1 is given in Algorithm 2. Other algorithms mentioned in Section 3.4: the standard binary segmentation [Vostrikova, 1981] is given in Algorithm 3, the construction of seeded intervals is given in Algorithm 4, and the seeded binary segmentation is given in Algorithm 5.

**Algorithm 2:** Pruning: CCP localization given candidates

**input:** data $(\mathbf{X}, \mathbf{Y})$, CCP candidates $\{k_1, \ldots, k_L\}$ with $k_i < k_j$ if $i < j$, and tests $(\phi_I)_{I \in \mathcal{I}}$ for $(\mathcal{H}_0^I)_{I \in \mathcal{I}}$

$k_0 \leftarrow 1$; $k_{L+1} \leftarrow n + 1$; $\widehat{\mathcal{T}} \leftarrow \emptyset$
**for** $\ell \in \{1, \ldots, L\}$ **do**
  $I \leftarrow \{k_{\ell-1}, \ldots, k_{\ell+1} - 1\}$
  **if** $\phi_I(\mathbf{X}_I, \mathbf{Y}_I) = 1$ **then**
    $\widehat{\mathcal{T}} \leftarrow \widehat{\mathcal{T}} \cup \{k_\ell\}$

**output:** $\widehat{\mathcal{T}}$

**Algorithm 3:** Binary segmentation

**input:** data $(\mathbf{X}, \mathbf{Y})$, tests $(\phi_I)_{I \in \mathcal{I}}$ for $(\mathcal{H}_0^I)_{I \in \mathcal{I}}$
*tuning*: a minimal segmentation length $s$

**function** `BinSeg`$(\mathbf{X}, \mathbf{Y}, b, e, s)$**:**
  Let $I \coloneqq \{b, \ldots, e\}$
  **if** $|I| > h$ *and* $\phi_I = 1$ **then**
    $k \coloneqq \arg\min_{i \in I} \widehat{\mathcal{L}}_{I,s}(i)$
    $G \coloneqq$ `BinSeg`$(\mathbf{X}, \mathbf{Y}, b, k - 1, s)$
    $D \coloneqq$ `BinSeg`$(\mathbf{X}, \mathbf{Y}, k, e, s)$
    **return** $G \cup \{d\} \cup D$
  **else**
    **return** $\emptyset$
Let $\widehat{\mathcal{D}} \coloneqq$ `BinSeg`$(\mathbf{X}, \mathbf{Y}, 1, n, s)$.

**output:** $\widehat{\mathcal{D}}$

## 3.C. Additional numerical experiments and experiment details

### 3.C.1 Additional numerical experiments

Further details on the following experiments can be found in Appendix 3.C.2.

**Algorithm 4:** Seeded intervals

**input:** Total number of time points $n$

*tuning*: a decay number $a \in [1/2, 1)$, and a minimal segmentation length $s$

Let $\mathfrak{B}_1 \coloneqq \{\{1, \ldots, n\}\}$ and $L \coloneqq \lfloor 1 + \log_{\frac{1}{a}} \frac{n}{s} \rfloor$ `// L is the number of levels`

**for** $\ell \in \{2, \ldots, L\}$ **do**

  Let $h_\ell \coloneqq na^{\ell-1}$, $s_\ell \coloneqq \frac{n-h_\ell}{q_\ell - 1}$, and $q_\ell \coloneqq 2\lceil \frac{1}{a^{\ell-1}} \rceil$

  **for** $j \in \{1, \ldots, q_\ell\}$ **do**

    $I_j \coloneqq \left\{ \lfloor (j-1)s_\ell \rfloor + 1, \ldots, \lceil (j-1)s_\ell + h_\ell \rceil \right\}$

  $\mathcal{B}_\ell \coloneqq \left\{ I_1, \ldots, I_{q_\ell} \right\}$ `// B_ℓ denotes level ℓ, which is a set of intervals`

**output:** $L$ levels of intervals $\mathcal{B}_1, \cdots, \mathcal{B}_L$

**Algorithm 5:** Seeded binary segmentation (with narrowest-over-threshold)

**input:** data $(\mathbf{X}, \mathbf{Y})$, tests $(\phi_I)_{I \in \mathcal{I}}$ for $(\mathcal{H}_0^I)_{I \in \mathcal{I}}$, and levels of intervals $\mathcal{B}_1, \ldots, \mathcal{B}_L$ obtained from Algorithm 4

Let $\widehat{\mathcal{T}} \coloneqq \emptyset$ and $i \coloneqq 0$

**while** $i < L$ **do**

  **if** *there exists* $I \in \mathcal{B}_{L-i}$ *such that* $\phi_I = 1$ **then**

    Let $I$ be the interval in $\mathcal{B}_{L-i}$ that has the smallest p-value

    Let $k \coloneqq \arg\min_{i \in I} \widehat{\mathcal{L}}_{I,s}(i)$ and $\widehat{\mathcal{T}} \coloneqq \widehat{\mathcal{T}} \cup \{k\}$

    Update $\mathcal{B}_1, \ldots, \mathcal{B}_L$ by removing all intervals in all $\mathcal{B}_1, \ldots, \mathcal{B}_L$ that contain $k$

  **else**

    Remove $\mathcal{B}_i$

  $i \coloneqq i + 1$

**output:** $\widehat{\mathcal{T}}$

**Experiment 4: Pruning when the candidate set is a subset of the true CCPs** For a total number of time points $n$, two CCPs are fixed at $\lceil 0.25n + 1 \rceil$ and $\lceil 0.75n + 1 \rceil$. We illustrate the performance of pruning described in Section 3.4.1 when the candidate set contains only one of the two CCPs. Figure 3.C.1 shows the power of detecting all CCPs contained in the candidate set when the candidate set contains only the first CCP ($k_1$), only the second CCP ($k_2$), or contains both CCPs ($k_1, k_2$). In this setting, we observe that when the candidate set contains only one of the two true CCPs, the detection of this CCP is not affected for a large enough sample size.

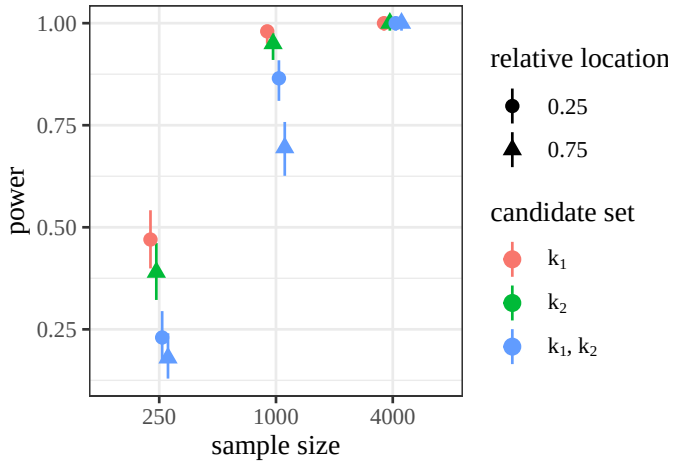



Figure 3.C.1: Power of detecting all CCPs contained in the candidate set when the candidate set contains only one true CCP or both of the true CCPs.

**Experiment 5: Pruning when the candidate deviates from the true CCP** For $n = 4000$, one true CCP is fixed at $0.5n$. We let the candidate equal $(0.5 + \delta)n$ for different $\delta$. Figure 3.C.2 shows (with the marker '$\times$') the percentage of repetitions for which the method (incorrectly for $\delta \neq 0$) outputs the candidate as a CCP, i.e., the test that the candidate is not a CCP is rejected (with $\alpha = 0.05$). If the method does output the candidate as a CCP, we test whether there exists an invariant set over the sub-intervals to the left and right of the candidate. The percentages of repetitions where the test is rejected in at least one of the sub-intervals versus not rejected in both sub-intervals given different $\delta$'s are also shown with bar charts in Figure 3.C.2. Even though an inaccurate candidate can be output as an estimated CCP, it is possible to invalidate the output by testing for the existence of invariant sets over the sub-intervals divided by the candidate.

## 3.C.2 Experiment details

For each of the experiments we first choose the parameters of the SCMs (3.5.8) for all $i \in \{1, \ldots, n\}$, and then generate one dataset accordingly for each of the 200 repetitions. As before, the parameters are chosen such that changes in the causal parameters are CCPs and changes in the non-causal parameters are NCCPs (we have verified this

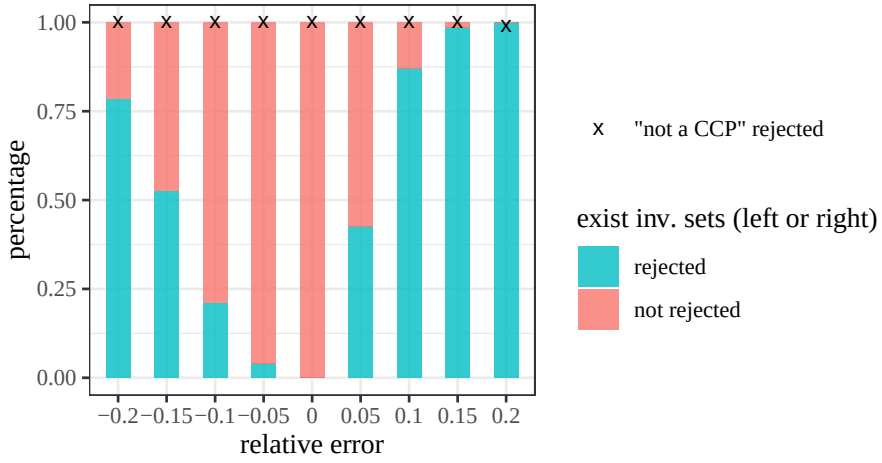



Figure 3.C.2: When the candidate deviates from the true CCP by a relative error $\delta \in \{-0.2, -0.15, \ldots, 0.2\}$, the percentages of cases where 'the candidate is not a CCP' is rejected, are almost all at 1 (when $\delta = 0$ the rejection is correct while when $\delta \neq 0$ the rejection is incorrect). However, we can test whether there is an invariant set in the two sub-intervals to the left or right of the candidate that is output as a CCP. If we reject that there is an invariant set in one of the two sub-intervals to the left or right of this candidate, it indicates that classifying the candidate as a CCP may be incorrect.

using straight-forward computations). The specific values of these parameters for each experiment in Section 3.5.1.2 are given below.

- *CCP detection in Section 3.5.1.1*: Table 3.C.1.
- *CCP localization Experiment 1 in Section 3.5.1.2*: Table 3.C.2.
- *CCP localization Experiment 2 in Section 3.5.1.2*: Table 3.C.3.
- *CCP localization Experiment 3 in Section 3.5.1.2*: Table 3.C.4.
- *CCP localization Experiment 4 in Appendix 3.C.1*: Table 3.C.5.
- *CCP localization Experiment 5 in Appendix 3.C.1*: Table 3.C.6.

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1, \ldots, n\nu$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $n\nu + 1, \ldots, n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |

Table 3.C.1: Parameter values of CCP detection experiment in Section 3.5.1.1 where $\nu = \dfrac{\tau - 1}{n}$ is the relative location of the single CCP, $\nu \in \{0.1, \ldots, 0.9\}$.

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1,\dots,0.25n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $0.25n+1,\dots,0.5n$ | 1.50 | 0.50 | 0.50 | 1.50 | 1.00 | 0.71 | 0.71 | 1.22 | 1.22 | 1.00 | 1.50 | 1.50 | 0.50 | 1.00 | 1.00 |
| $0.5n+1,\dots,0.75n$ | 1.50 | 0.50 | 0.50 | 1.50 | 0.50 | 0.71 | 0.71 | 1.22 | 1.22 | 1.22 | 1.50 | 1.50 | 0.50 | 1.50 | 0.50 |
| $0.75n+1,\dots,n$ | 0.75 | 0.75 | 0.25 | 0.75 | 0.50 | 0.50 | 0.50 | 1.50 | 0.87 | 1.22 | 2.25 | 0.75 | 0.25 | 1.50 | 0.50 |

Table 3.C.2: Parameter values of CCP localization Experiment 1 in Section 3.5.1.2 where $0.5n$ is the location of the single CCP, and $0.25n$ and $0.75n$ are the locations of two NCCPs. The changes at the CCP and the NCCPs are such that each causal (respectively, non-causal) parameter either increase or decrease (with probability 0.5) by 50% of its value at the previous time point (for $\sigma_i^j$ where $j \in \{1,2,3,Y\}$, the changes are 50% in their squared values).

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1,\dots,200$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $201,\dots,500$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 1.00 | 1.00 | 1.50 | 1.50 |
| $501,\dots,1500$ | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 | 1.22 | 1.22 | 1.22 | 1.22 | 1.22 | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 |
| $1501,\dots,2000$ | 2.25 | 2.25 | 2.25 | 2.25 | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 | 1.22 | 2.25 | 2.25 | 2.25 | 1.50 | 1.50 |

Table 3.C.3: An example of the parameter values of CCP localization experiment in Section 3.5.1.2 with $\nu = 0.1$. In this experiment, $n = 2000$, two fixed NCCPs are located at 501 and 1501, and one CCP is located at $\tau = n\nu + 1$. The changes at the CCP and the NCCPs are such that each causal parameter (respectively, non-causal) increases by 50% of its value at the previous time point (for $\sigma_i^j$ where $j \in \{1,2,3,Y\}$, the changes are 50% in their squared values). The parameter values for $\nu \in \{0.2,\dots,0.9\}$ are constructed in the same way.

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1,\dots,0.25n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $0.25n+1,\dots,0.5n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| $0.5n+1,\dots,0.75n$ | 1.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 | 1.00 | 1.00 | 0.00 |
| $0.75n+1,\dots,n$ | 1.00 | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 | 1.00 | 2.00 | 1.00 | 0.00 | 1.00 |

Table 3.C.4: Parameter values of CCP localization experiment Experiment 1 in Section 3.5.1.2 where $0.5n$ is the location of the single CCP, and $0.25n$ and $0.75n$ are the locations of two NCCPs.

### 3.C.3 Real data details and additional results

A description of all variables in the Beijing air quality example is given in Table 3.C.7. We also include the variable description of the monetary policy example from Pfister et al. [2019] in Table 3.C.8 for completeness. All variables are scaled to have mean 0 and variance 1 before applying any change point detection and localization methods. The estimated RCP and MCP dates in Section 3.5.2.1 (Table 3.5.1) are given in Table 3.C.9.

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1, \dots, 0.25n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $0.25n+1, \dots, 0.75n$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 1.00 | 1.00 | 1.50 | 1.50 |
| $0.75n+1, \dots, n$ | 1.00 | 1.00 | 1.00 | 1.00 | 2.25 | 1.00 | 1.00 | 1.00 | 1.00 | 1.50 | 1.00 | 1.00 | 1.00 | 2.25 | 2.25 |

Table 3.C.5: Parameter values of CCP localization experiment Experiment 4 in Appendix 3.C.1 where $0.25n$ and $0.75n$ are the locations of two CCPs. The changes at the CCP are such that each causal parameter increases by 50% of its value at the previous time point (for $\sigma_i^Y$ the changes are 50% in its squared value).

| $i$ | $\mu_i^1$ | $\mu_i^2$ | $\mu_i^3$ | $\mu_i^4$ | $\mu_i^Y$ | $\sigma_i^1$ | $\sigma_i^2$ | $\sigma_i^3$ | $\sigma_i^4$ | $\sigma_i^Y$ | $a_i^{12}$ | $a_i^{53}$ | $a_i^{43}$ | $b_i^{15}$ | $b_i^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1, \dots, 2000$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $2000, \dots, 4000$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 | 1.00 | 1.00 | 1.00 | 1.50 | 1.50 |

Table 3.C.6: Parameter values of CCP localization experiment Experiment 5 in Appendix 3.C.1 where 2000 is the location of the single CCP. The changes at the CCP are such that each causal parameter increases by 50% of its value at the previous time point (for $\sigma_i^Y$ the change is 50% in its squared value).

| Variable | Description |
|---|---|
| $Y$ | log daily average of PM2.5 concentration |
| $X^1$ | log daily average of SO2 concentration |
| $X^2$ | log daily average of NO2 concentration |
| $X^3$ | log daily average of CO concentration |
| $X^4$ | daily average drew point temperature (DEWP) |
| $X^5$ | daily average wind speed (WSPM) |
| $X^6$ | heating indicator (1 if during heating period and 0 otherwise) |

Table 3.C.7: Variable description of the Beijing air quality dataset.

| Variable | Description |
|---|---|
| $Y$ | log returns of end of month exchange rate Euro to Swiss Francs |
| $X^1$ | change in average call money rate (no log transform as part of the values are negative) |
| $X^2$ | log returns of end of month proportion of foreign currency investments from total assets on the balance sheet of the SNB |
| $X^3$ | log returns of end of month proportion of reserve positions at International Monetary Fund (IMF) from total assets on the balance sheet of the SNB |
| $X^4$ | log returns of end of month proportion of monetary assistance loans from total assets on the balance sheet of the SNB |
| $X^5$ | log returns of end of month proportion of Swiss Franc securities from total assets on the balance sheet of the SNB |
| $X^6$ | log returns of end of month proportion of remaining assets from total assets on the balance sheet of the SNB |
| $X^7$ | log returns of Swiss GDP (in Euro) resulting from interpolation of quarterly (seasonally adjusted) data and adjusted using the monthly average exchange rate |
| $X^8$ | log returns of Euro zone GDP resulting from an interpolation of quarterly (seasonally adjusted) GDP data |
| $X^9$ | inflation rate for Switzerland computed from the monthly consumer price index (CPI) |

Table 3.C.8: Variable description of the monetary policy dataset from Pfister et al. [2019] Table 2.

| Method | Number of lags | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| RCP | 2014-05-24 | 2014-05-23 | 2014-05-22 | 2014-05-24 | 2014-05-23 | 2014-05-21 | 2014-05-24 |
| | 2014-11-11 | 2014-11-15 | 2014-11-11 | 2014-11-10 | 2014-11-12 | 2014-11-08 | 2014-11-07 |
| | 2015-03-03 | 2015-03-03 | 2015-02-28 | 2015-02-28 | 2015-02-28 | 2015-02-26 | 2015-02-25 |
| | 2015-06-20 | 2015-06-19 | 2015-06-16 | 2015-06-16 | 2015-06-16 | 2015-06-14 | 2015-06-13 |
| | 2015-10-06 | 2015-10-05 | 2015-10-02 | 2015-10-02 | 2015-10-02 | 2015-09-30 | 2015-09-29 |
| | 2016-01-27 | 2016-01-26 | 2016-01-18 | 2016-01-19 | 2016-01-18 | 2016-01-17 | 2016-01-16 |
| | 2016-05-20 | 2016-05-19 | 2016-05-18 | 2016-05-17 | 2016-05-16 | 2016-05-15 | 2016-05-14 |
| | 2016-09-11 | 2016-09-10 | 2016-09-09 | 2016-09-08 | 2016-09-07 | 2016-09-07 | 2016-09-06 |
| MCP | 2014-02-05, 2014-02-16, 2014-03-15, 2014-03-21, 2014-07-01, 2014-09-29, 2014-10-04, 2014-10-06, 2014-11-11, 2014-11-23, 2014-12-18, 2015-01-19, 2015-03-01, 2015-04-24, 2015-06-13, 2015-08-13, 2015-09-06, 2015-09-26, 2015-09-30, 2015-10-05, 2015-11-01, 2015-11-08, 2015-11-19, 2015-11-24, 2015-11-27, 2015-12-07, 2015-12-11, 2015-12-27,2016-01-31, 2016-02-05, 2016-02-22, 2016-02-28, 2016-03-04, 2016-03-15, 2016-03-27, 2016-08-17, 2016-09-05, 2016-10-13, 2016-10-25, 2016-11-13, 2016-11-16, 2016-12-08, 2016-12-14 | | | | | | |

Table 3.C.9: Estimated dates of RCPs and MCPs in the Beijing air quality application based on the meteorological-variable-only model. The MCPs are estimated by WBS using random intervals. With different random seeds, we obtain similar estimates. The RCPs are estimated by BP, with the number of RCPs selected by minimizing the residual sum of squares.

## 3.D. Proofs

### 3.D.1 Proof of Proposition 3.2.6

*Proof.* We show both directions separately.

($\Rightarrow$) Assume $k$ is a CCP and fix an arbitrary $S \in \mathcal{S}$. Then, since $k$ is a CCP it holds that either $\beta_k^{\mathrm{OLS}}(S) \neq \beta_{k-1}^{\mathrm{OLS}}(S)$ or $\epsilon_k(S) \overset{d}{\neq} \epsilon_{k-1}(S)$. This implies, that $S$ is not a $\{k-1,k\}$-invariant set. Since $S$ was arbitrary this implies the result.

($\Leftarrow$) Assume there does not exist a $\{k-1,k\}$-invariant set $S \in \mathcal{S}$. Then, it holds for all $S \in \mathcal{S}$ that either $\beta_k^{\mathrm{OLS}}(S) \neq \beta_{k-1}^{\mathrm{OLS}}(S)$ or $\epsilon_k(S) \overset{d}{\neq} \epsilon_{k-1}(S)$, which implies that $k$ is a CCP.

This completes the proof of Proposition 3.2.6. □

### 3.D.2 Proof of Proposition 3.2.7

*Proof.* We begin by connecting the causal coefficient and residual with the corresponding population OLS coefficient and residual given the parents of $Y$. By the definitions of the causal and population OLS coefficients it holds for all $i \in \{k-1,k\}$ and all $j \in \{1,\dots,d+1\} \setminus \mathrm{PA}(Y_i)$, that $(\beta_i^{\mathrm{OLS}}(\mathrm{PA}(Y_i)))^j = \beta_i^j = 0$ and

$$\begin{aligned}
\Big(\beta_i^{\mathrm{OLS}}(\mathrm{PA}(Y_i))\Big)^{\mathrm{PA}(Y_i)} &= \mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}(X_i^{\mathrm{PA}(Y_i)})^\top\Big]^{-1}\mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}Y_i\Big] \\
&= \mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}(X_i^{\mathrm{PA}(Y_i)})^\top\Big]^{-1}\mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}(X_i^{\mathrm{PA}(Y_i)})^\top\beta_i^{\mathrm{PA}(Y_i)} \\
&\qquad + X_i^{\mathrm{PA}(Y_i)}g_i(H_i,\epsilon_i^Y)\Big] \\
&= \beta_i^{\mathrm{PA}(Y_i)} + \mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}(X_i^{\mathrm{PA}(Y_i)})^\top\Big]^{-1}\mathbb{E}\Big[X_i^{\mathrm{PA}(Y_i)}g_i(H_i,\epsilon_i^Y)\Big] \\
&= \beta_i^{\mathrm{PA}(Y_i)}, \qquad (3.D.1)
\end{aligned}$$

where the last equality follows from the assumption that for all $i \in \{k-1,k\}$, $\mathbb{E}[X_i^{\mathrm{PA}(Y_i)}g_i(H_i,\epsilon_i^Y)] = 0$. Thus, we get for all $i \in \{k-1,k\}$ that

$$\beta_i^{\mathrm{OLS}}(\mathrm{PA}(Y_i)) = \beta_i. \qquad (3.D.2)$$

Moreover, using this result and the definition of the population OLS residual given $\mathrm{PA}(Y_i)$ we also obtain that

$$\begin{aligned}
\epsilon_i(\mathrm{PA}(Y_i)) &= Y_i - X_i^\top\beta_i^{\mathrm{OLS}}(\mathrm{PA}(Y_i)) \\
&= Y_i - X_i^\top\beta_i \\
&= g_i(H_i,\epsilon_i^Y). \qquad (3.D.3)
\end{aligned}$$

Now, we can prove the result. To this end, assume $k \in \{2,\dots,n\}$ satisfies that $\beta_k = \beta_{k-1}$ and $g_k(H_k,\epsilon_k^Y) \overset{d}{=} g_{k-1}(H_{k-1},\epsilon_{k-1}^Y)$. This implies that $\mathrm{PA}(Y_{k-1}) = \mathrm{PA}(Y_k)$ and hence, using (3.D.2) and (3.D.3), the set $\mathrm{PA}(Y_k)$ is $\{k-1,k\}$-invariant. Therefore, by Proposition 3.2.6, $k$ is not a CCP. This completes the proof of Proposition 3.2.7. □

### 3.D.3 Proof of Proposition 3.3.1

*Proof.* Suppose $(\mathbf{X}_I, \mathbf{Y}_I) \sim P \in \mathcal{H}_0^I$ and let $S \in \mathcal{S}$ be s.t. $S$ is $I$-invariant. We then have

$$\mathbb{P}_P(\phi_I = 1) \leq \mathbb{P}_P(\phi_I^S = 1) \leq \alpha.$$

□

### 3.D.4 Proof of Proposition 3.4.1

*Proof.* For (i), assume that for all $k \in \mathcal{K}$ the test $\phi_{I_k}$ is level $\alpha$. Then, using a union bound we get

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{T}} \subseteq \mathcal{T}) &= 1 - \mathbb{P}(\exists k \in \widehat{\mathcal{T}} : k \notin \mathcal{T}) \\
&\geq 1 - \sum_{k \in \mathcal{K} \setminus \mathcal{T}} \mathbb{P}(k \in \widehat{\mathcal{T}}) \\
&\geq 1 - \sum_{k \in \mathcal{K} \setminus \mathcal{T}} \mathbb{P}(\phi_{I_k} = 1) \\
&\geq 1 - (|\mathcal{K}| - |\mathcal{T}|) \cdot \alpha,
\end{aligned}$$

where in the last step we used that $\mathcal{H}_0^{I_k}$ is true for all $k \in \mathcal{K} \setminus \mathcal{T}$.

Similarly, for (ii), assume that for all $\ell \in \{1, \ldots, L\}$ with $k_\ell \in \mathcal{T}$ it holds that $\mathbb{P}(\phi_{I_{k_\ell}} = 1) \leq \beta$, then

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{T}} \supseteq \mathcal{T}) &= 1 - \mathbb{P}(\exists k \in \mathcal{T} : k \notin \widehat{\mathcal{T}}) \\
&\geq 1 - \sum_{k \in \mathcal{T}} \mathbb{P}(k \notin \widehat{\mathcal{T}}) \\
&= 1 - \sum_{k \in \mathcal{T}} \mathbb{P}(\phi_{I_k} = 0) \\
&\geq 1 - |\mathcal{T}| \cdot (1 - \beta).
\end{aligned}$$

This concludes the proof. □

### 3.D.5 Proof of Proposition 3.4.3

*Proof.* Suppose there is no CCP in $I$. Then, there exists a set $S \in \mathcal{S}$, a parameter $\beta \in \mathbb{R}^{d+1}$ and a distribution $F_\epsilon$ such that for all $i \in I$ it holds that $\beta_i^{\mathrm{OLS}}(S) = \beta$ and $\epsilon_i(S) \overset{\text{iid}}{\sim} F_\epsilon$. Moreover, by Lemma 1 it holds for all $J \subseteq I$ and all $i \in I$ that $\epsilon_i^J(S) \overset{\text{iid}}{\sim} F_\epsilon$. Hence, we get for all $J, J' \subseteq I$ that

$$\begin{aligned}
V_{J',J}(S) &= \frac{1}{|J'|} \sum_{\ell \in J'} \mathbb{E}\left[\epsilon_\ell^J(S)^2\right] \\
&= \frac{1}{|J'|} \sum_{\ell \in J'} \mathbb{E}_{\nu \sim F_\epsilon}\left[\nu^2\right] \\
&= \mathbb{E}_{\nu \sim F_\epsilon}\left[\nu^2\right].
\end{aligned}$$

Therefore, for $s \in \mathbb{N}$, we get

$$\begin{aligned}
\mathcal{C}_s(I) &= \min_{\tilde{S} \in \mathcal{S}} \sum_{r=1}^{m_s(I)} \left( V_{P_r^c(I),P_r(I)}(\tilde{S}) - V_{P_r(I),P_r(I)}(\tilde{S}) \right)^2 \\
&\leq \sum_{r=1}^{m_s(I)} \left( V_{P_r^c(I),P_r(I)}(S) - V_{P_r(I),P_r(I)}(S) \right)^2 \\
&= \sum_{r=1}^{m_s(I)} \left( \mathbb{E}_{\nu \sim F_\epsilon}\left[\nu^2\right] - \mathbb{E}_{\nu \sim F_\epsilon}\left[\nu^2\right] \right) \\
&= 0.
\end{aligned}$$

Now, since $\mathcal{C}_s(I) \geq 0$, we have that $\mathcal{C}_s(I) = 0$. This completes the proof of Proposition 3.4.3.

□

## 3.E. Auxiliary results

**Lemma 1.** *Let $I \in \mathcal{I}$ and $S \in \mathcal{S}$ is an $I$-invariant set. Then, for all $J \subseteq I$ and all $i \in I$, it holds that $\beta_J^{OLS}(S) = \beta_i^{OLS}(S)$, and that there exists a distribution $F_\epsilon$ such that $\epsilon_i^J(S) = \epsilon_i(S) \overset{\text{iid}}{\sim} F_\epsilon$.*

*Proof.* Since $S$ is $I$-invariant, there exists $\beta \in \mathbb{R}^{d+1}$ and distribution $F_\epsilon$ on $\mathbb{R}$ such that for all $i \in I$ it holds that $\beta_i^{\text{OLS}}(S) = \beta$ and $\epsilon_i(S) \overset{\text{iid}}{\sim} F_\epsilon$. Moreover, since the population OLS coefficient satisfies for all $i \in I$ that $\mathbb{E}\left[X_i^S Y_i\right] = \mathbb{E}\left[X_i^S (X_i^S)^\top\right] \beta_i^{\text{OLS}}(S)^S$, it immediately follows for all $J \subseteq I$ that

$$\begin{aligned}
\left(\beta_J^{\text{OLS}}(S)\right)^S &= \Big[\sum_{\ell \in J} \mathbb{E}[X_\ell^S (X_\ell^S)^\top]\Big]^{-1} \sum_{\ell \in J} \mathbb{E}[X_\ell^S Y_\ell] \\
&= \Big[\sum_{\ell \in J} \mathbb{E}[X_\ell^S (X_\ell^S)^\top]\Big]^{-1} \sum_{\ell \in J} \mathbb{E}\left[X_\ell^S (X_\ell^S)^\top\right] \beta_\ell^{\text{OLS}}(S)^S \\
&= \Big[\sum_{\ell \in J} \mathbb{E}[X_\ell^S (X_\ell^S)^\top]\Big]^{-1} \sum_{\ell \in J} \mathbb{E}\left[X_\ell^S (X_\ell^S)^\top\right] \beta^S \\
&= \beta^S.
\end{aligned}$$

Since, for all $j \in \{1, \ldots, d+1\} \setminus S$ it also holds that $\left(\beta_J^{\text{OLS}}(S)\right)^j = \beta^j = 0$, we get that $\beta_J^{\text{OLS}}(S) = \beta$. Moreover, this further implies for all $i \in I$ that

$$\epsilon_i^J(S) = Y_i - X_i^\top \beta_J^{\text{OLS}}(S) = Y_i - X_i^\top \beta = Y_i - X_i^\top \beta_i^{\text{OLS}}(S) = \epsilon_i(S) \overset{\text{iid}}{\sim} F_\epsilon,$$

this completes the proof of Lemma 1. □

## 3.F. Chow test

Here we review the Chow test [Chow, 1960] where we adapted the setup and notations to this paper.

Let $I = \{t_1, \ldots, t_l\} \in \mathcal{I}$ where $l > d$, and consider an arbitrary $k \in I$, $I_k^1 \coloneqq \{t_1, \ldots, k\}$, and $I_k^2 \coloneqq \{k+1, \ldots, t_l\}$ be the two non-overlapping subsets splitting $I$ at $k$. Denote $l_1 \coloneqq |I_k^1|$ and $l_2 \coloneqq l - l_1$.

Assume that $l_1 > d$, and for all $m \in \{1, 2\}$ and $i \in I_k^m$,

$$Y_i = X_i^S \beta_m + \epsilon_i \quad \text{and} \quad \mathbb{E}[\epsilon_i \mid X_i^S] = 0 \tag{3.F.1}$$

with $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Then, the null hypothesis

$$\mathcal{H}_0^S(I, k) : \beta_1 = \beta_2 \tag{3.F.2}$$

holds if (3.3.6) holds. The Chow test [Chow, 1960] described below can be used for testing (3.F.2).

**Proposition 1** (Chow test)**.** *Let* $\hat{\beta}_{I_k^1} \coloneqq (\mathbf{X}_{I_k^1}^\top \mathbf{X}_{I_k^1})^{-1} \mathbf{X}_{I_k^1} \mathbf{Y}_{I_k^1}$, $\hat{\beta}_{I_k^2} \coloneqq (\mathbf{X}_{I_k^2}^\top \mathbf{X}_{I_k^2})^{-1} \mathbf{X}_{I_k^2} \mathbf{Y}_{I_k^2}$, *and* $\hat{\beta}_I \coloneqq (\mathbf{X}_I^\top \mathbf{X}_I)^{-1} \mathbf{X}_I \mathbf{Y}_I$. *Denote the residuals by* $R_{I_k^1} \coloneqq \mathbf{Y}_{I_k^1} - \mathbf{X}_{I_k^1} \hat{\beta}_{I_k^1}$ *and* $R_{I_k^2} \coloneqq \mathbf{Y}_{I_k^2} - \mathbf{X}_{I_k^2} \hat{\beta}_{I_k^2}$. *Then, under the null hypothesis* $\mathcal{H}_0^S(I, k)$ *the following two statements hold;*

- *if* $l_2 > d$,

$$\frac{||\mathbf{X}_{I_k^1} \hat{\beta}_{I_k^1} - \mathbf{X}_{I_k^2} \hat{\beta}_I||^2 + ||\mathbf{X}_{I_k^2} \hat{\beta}_{I_k^2} - \mathbf{X}_{I_k^2} \hat{\beta}_I||^2}{||R_{I_k^1}||^2 + ||R_{I_k^2}||^2} \cdot \frac{l - 2d}{d} \sim F(d, l - 2d) \tag{3.F.3}$$

- *if* $l_2 \leq d$,

$$\begin{aligned} &\frac{(\mathbf{Y}_{I_k^2} - \mathbf{X}_{I_k^2} \hat{\beta}_{I_k^1})^\top [\mathbf{I}_{l_2} + \mathbf{X}_{I_k^2} (\mathbf{X}_{I_k^1}^\top \mathbf{X}_{I_k^1})^{-1} \mathbf{X}_{I_k^2}^\top]^{-1} (\mathbf{Y}_{I_k^2} - \mathbf{X}_{I_k^2} \hat{\beta}_{I_k^1})}{||R_{I_k^1}||^2} \cdot \frac{l_1 - d}{l_2} \\ =&\frac{||\mathbf{X}_{I_k^1} \hat{\beta}_{I_k^1} - \mathbf{X}_{I_k^2} \hat{\beta}_I||^2 + ||\mathbf{Y}_{I_k^2} - \mathbf{X}_{I_k^2} \hat{\beta}_I||^2}{||R_{I_k^1}||^2} \cdot \frac{l_1 - d}{l_2} \\ \sim& F(l_2, l_1 - d), \end{aligned} \tag{3.F.4}$$

  *where* $\mathbf{I}_l$ *denotes the identity matrix of dimension* $l$.

*Proof.* See Chow [1960]. □

# 4 Sparse causal effect estimation using two-sample summary statistics in the presence of unmeasured confounding

Shimeng Huang, Niklas Pfister, and Jack Bowden

**Abstract**

Observational genome-wide association studies are now widely used for causal inference in genetic epidemiology. To maintain privacy, such data is often only publicly available as summary statistics, and often studies for the endogenous covariates and the outcome are available separately. This has necessitated methods tailored to two-sample summary statistics. Current state-of-the-art methods modify linear instrumental variable (IV) regression—with genetic variants as instruments—to account for unmeasured confounding. However, since the endogenous covariates can be high dimensional, standard IV assumptions are generally insufficient to identify all causal effects simultaneously. We ensure identifiability by assuming the causal effects are sparse and propose a sparse causal effect two-sample IV estimator, `spaceTSIV`, adapting the `spaceIV` estimator by Pfister and Peters [2022] for two-sample summary statistics. We provide two methods, based on L0- and L1-penalization, respectively. We prove identifiability of the sparse causal effects in the two-sample setting and consistency of `spaceTSIV`. The performance of `spaceTSIV` is compared with existing two-sample IV methods in simulations. Finally, we showcase our methods using real proteomic and gene-expression data for drug-target discovery.

## 4.1 Introduction

The use of observational data to study the causal effects of covariate interventions on an outcome has seen a surge in popularity in many scientific areas. A primary example is genetic epidemiology, where a common research topic is to study the causal effects of genetically predictive phenotypic traits, such as a person's body mass index or their low density lipoprotein cholesterol, on downstream disease outcomes. This is often based on Mendelian randomization (MR)—that is, instrumental variable estimation (IV) with genetic variants being the instruments—to account for unmeasured confounding between

the endogenous covariates and the outcome. However, due to privacy concerns, access to individual-level genetic data is highly regulated. To both preserve privacy and enable data sharing, public data repositories of genetic summary statistics are made available by various international genome-wide association study (GWAS) consortia. These summary statistics usually contain marginal effect estimates of single nucleotide polymorphisms (SNPs) on the phenotypic traits and disease outcomes, along with their standard errors, which are often themselves obtained from two separate GWAS. This is referred to as the "two-sample summary statistics" setting. Zhao et al. [2019] discuss sufficient assumptions that enable consistent estimation under two-sample IV, specifically the homogeneity of the two samples. When the number of endogenous covariates under investigation is high dimensional or the instruments are highly correlated, a case in point being human gene expression phenotypes and genetic variants, there may be an insufficient number of strong and valid instruments to ensure the identifiability of the multivariable causal effects.

Lack of identifiability leads to poor estimation, or weak instrument bias. In the univariable two-sample summary statistics setting, Bowden et al. [2019] develop heuristic weak-instrument robust inference strategies based on heterogeneity statistic estimating equations. Under the same setting, Wang and Kang [2022] further clarify the connection between these approaches and summary statistics analogues of the Anderson-Rubin (AR) test statistic [Anderson and Rubin, 1949] and Limited Information Maximum Likelihood (LIML). Wang et al. [2021b] further extend weak-instrument robust models to the multivariable case. Another way to circumvent the weak instrument problem is to employ principal component analysis (PCA). Building on the work of Batool et al. [2022], Patel et al. [2024] show how many individually weak variants could be fashioned into PCA scores with improved instrument strength.

An alternative strategy to tackle the lack of identifiability is to introduce sparsity assumption on the causal effects. This is often a reasonable assumption in MR studies, as it is usually the case that many endogenous traits do not have direct causal effects on the outcome. Under the assumptions of independent instruments and the number of instruments is no less than the number of covariates, Grant and Burgess [2022] consider the use of L1 penalization on the causal effects in multivariable MR models where one covariate is of special interest but the others are allowed to be penalized. In related works, Rees et al. [2019], Zhao et al. [2020] and Grant and Burgess [2021] consider L1 penalization for individual instruments suspected to be invalid due to exclusion restriction violation, rather than penalization on the number of causal effects.

In the one sample individual-level data setting, the identifiability conditions for sparse causal effects have been studied by Pfister and Peters [2022], and they propose a sparse causal effect estimator, `spaceIV`. Tang et al. [2023] also consider sparse causal effect identification and estimation under assumptions on the sparsity level and propose a synthetic two-stage regularized regression approach.

We propose `spaceTSIV`, adapting the `spaceIV` estimator for two-sample summary statistics. We allow the IVs to be correlated by extending the adjustment method in Wang and Kang [2022]. Two specific approaches based on L0- and L1-penalization, respectively, are provided. We prove identifiability of the sparse causal effects and con-

sistency of `spaceTSIV` under the two-sample summary statistics setting. We evaluate the performance of `spaceTSIV` with simulated data and compare it with existing (non-sparse) methods that work with two-sample summary statistics. Finally, we showcase our methods using proteomic and gene-expression data within the context of a drug-target discovery analysis. Notation is summarized below and all proofs are provided in Supplementary Material 4.C.

**Notation 4.1.1.** *For all $k \in \mathbb{N}$, we define $[k] \coloneqq \{1, \dots, k\}$ and for all $\beta \in \mathbb{R}^d$, we denote by $\operatorname{supp}(\beta) \coloneqq \{j \in [d] : \beta^j \neq 0\}$ the set of non-zero components of $\beta$. For an arbitrary matrix $A \in \mathbb{R}^{n \times m}$, we denote for all $i \in [n]$ and $j \in [m]$, the $i$-th row of $A$ by $A_i$, the $j$-th column of $A$ by $A^j$, and the $ij$-th entry of $A$ by $A_i^j$. If $A$ is a square block matrix containing $k \times k$ square matrices of dimension $l \times l$, then $A^{[ij]}$ for all $i, j \in [k]$ denotes the $ij$-th block of $A$.*

## 4.2 Reduced form IV model and summary statistics

We start from the conventional one-sample individual-level data setting and assume we observe $n$ independently and identically distributed (iid) observations $\{(X_i, Y_i, Z_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m$, where $Y$ is a response variable, $X$ a vector of endogenous covariates, and $Z$ a vector of instruments. The IV model assumptions can then be expressed as a linear structural causal model (SCM) over these variables.[1] Formally, for all $i \in [n]$, we assume,

$$\begin{aligned} X_i &\coloneqq AZ_i + BX_i + g(H_i, \nu_i^X) \\ Y_i &\coloneqq X_i^\top \beta^* + h(H_i, \nu_i^Y), \end{aligned} \tag{4.2.1}$$

where $H_i \in \mathbb{R}^q$ is a vector of unobserved variables, $g$ and $h$ are arbitrary measurable functions, $Z_i$, $h(H_i, \nu_i^X)$, and $g(H_i, \nu_i^Y)$ have mean 0 and finite variance, and $\{Z_i, H_i, \nu_i^X, \nu_i^Y\}_{i=1}^n$ are jointly independent. The coefficient $\beta^* \in \mathbb{R}^d$ denotes the true causal effect of the covariates on the response, and the matrices $A \in \mathbb{R}^{d \times m}$ and $B \in \mathbb{R}^{d \times d}$ encode the other causal relations in the SCM, with $B$ being a strictly lower triangular matrix. The matrix $I_d - B$ is assumed to be invertible, where $I_d$ is the identity matrix of dimension $d$. Finally, we call the support of $\beta^*$ the parent set of $Y$ and denote it as $\operatorname{PA}(Y)$, that is, $\operatorname{PA}(Y) \coloneqq \operatorname{supp}(\beta^*)$. The SCM (4.2.1) can also be expressed in what is called its reduced form by only considering how the instruments affect the covariates and the response. Formally, for all $i \in [n]$ the reduced form is given by

$$\begin{aligned} X_i &\coloneqq Z_i^\top \Pi + u_i^X \\ Y_i &\coloneqq Z_i^\top \pi + u_i^Y, \end{aligned} \tag{4.2.2}$$

where $\Pi \coloneqq A^\top (I_d - B)^{-\top} \in \mathbb{R}^{m \times d}$, $\pi \coloneqq \Pi \beta^* \in \mathbb{R}^m$, $u_i^X \coloneqq g(H_i, \nu_i^X)^\top (I_d - B)^{-\top}$, and $u_i^Y \coloneqq (u_i^X)^\top \beta^* + h(H_i, \nu_i)$.

[1] The required assumptions can also be expressed via other causal models (e.g., potential outcomes). Not all causal implications of the model introduced here are strictly necessary, but to keep the presentation concise we avoid presenting the most general assumptions.

In this work, we assume that we do not directly observe the individual-level data and instead only have access to summary statistics of partially observed paired data from two independent samples $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$ of the SCM (4.2.1).

As discussed in Section 4.1, this is often the case in MR studies utilizing summary statistics from two GWAS, one contains the associations between genetic variants and endogenous traits (such as gene expression levels), and the other contains the associations between genetic variants and an outcome trait (such as a disease), are used to study the causal relationship between the endogenous traits and the outcome trait with genetic variants being the IVs.

There are two types of summary statistics that we focus on here. Firstly, the two-sample joint OLS summary statistics, which consist of estimates of the reduced form parameters in (4.2.2) and are formally defined as follows.

**Definition 4.2.1** (Two-sample joint OLS summary statistics)**.** Given two independent samples of observations $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$, the *two-sample joint OLS summary statistics* (joint summary statistics) are defined as the set of estimates

$$\mathcal{D}_{a,b}^{\text{joint}} := \left\{\widehat{\pi}, \widehat{\Sigma}_\pi, \widehat{\Pi}, \widehat{\Sigma}_\Pi\right\},$$

where $\widehat{\pi} := (\mathbf{Z}_a^\top \mathbf{Z}_a)^{-1}\mathbf{Z}_a^\top \mathbf{Y}_a \in \mathbb{R}^m$, $\widehat{\Sigma}_\pi := \widehat{\varepsilon}_a^\top \widehat{\varepsilon}_a (\mathbf{Z}_a^\top \mathbf{Z}_a)^{-1} \in \mathbb{R}^{m\times m}$ with $\widehat{\varepsilon}_a := \mathbf{Y}_a - \mathbf{Z}_a\widehat{\pi}$, $\widehat{\Pi} := (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1}\mathbf{Z}_b^\top \mathbf{X}_b \in \mathbb{R}^{m\times d}$, and $\widehat{\Sigma}_\Pi \in \mathbb{R}^{md\times md}$ consists of $d \times d$ blocks of dimension $m \times m$ defined for all $k, l \in [d]$ by $\widehat{\Sigma}_\Pi^{[kl]} := (\widehat{\varepsilon}_b^k)^\top \widehat{\varepsilon}_b^l (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1}$ with $\widehat{\varepsilon}_b^k := \mathbf{X}_b^k - \mathbf{Z}_b\widehat{\Pi}^k$. ♣

Secondly, the two-sample marginal OLS summary statistics, which instead of capturing the joint effects described by the parameters in (4.2.2), only contain marginal univariate effects.

**Definition 4.2.2** (Two-sample marginal OLS summary statistics)**.** Given two independent samples of observations $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$, the *two-sample marginal OLS summary statistics* (marginal summary statistics) are defined as the set of estimates

$$\mathcal{D}_{a,b}^{\text{marginal}} := \left\{\widehat{\eta}, \widehat{\sigma}_\eta^2, \widehat{H}, \widehat{\sigma}_H^2, \widehat{M}_{Z_a}, \widehat{M}_{Z_b}, \widehat{M}_X\right\},$$

where $\widehat{\eta} \in \mathbb{R}^m$, $\widehat{\sigma}_\eta^2 \in \mathbb{R}^m$, $\widehat{H} \in \mathbb{R}^{m\times d}$, $\widehat{\sigma}_H^2 \in \mathbb{R}^{m\times d}$, and for all $j \in [m]$ and all $k \in [d]$, $\widehat{\eta}_j := (\mathbf{Z}_a^j)^\top \mathbf{Y}_a / (\mathbf{Z}_a^j)^\top \mathbf{Z}_a^j$, $\widehat{\sigma}_{\eta,j}^2 := (\widehat{\varepsilon}_a^j)^\top \widehat{\varepsilon}_a^j / \left((\mathbf{Z}_a^j)^\top \mathbf{Z}_a^j\right)$ with $\widehat{\varepsilon}_a^j := \mathbf{Y}_a - \widehat{\eta}_j \mathbf{Z}_a^j$, $\widehat{H}_j^k := (\mathbf{Z}_b^j)^\top \mathbf{X}_b^k / (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j$, and $(\widehat{\sigma}_{H,j}^k)^2 := (\widehat{\varepsilon}_{bj}^k)^\top \widehat{\varepsilon}_{bj}^k / \left((\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j\right)$ with $\widehat{\varepsilon}_{bj}^k := \mathbf{X}_b^k - \widehat{H}_j^k \mathbf{Z}_b^j$. For both $s \in \{a, b\}$, let $D_{Z_s}$ be the diagonal matrix containing the diagonal elements of $\mathbf{Z}_s^\top \mathbf{Z}_s$, then $\widehat{M}_{Z_s} := D_{Z_s}^{-1/2} \mathbf{Z}_s^\top \mathbf{Z}_s D_{Z_s}^{-1/2} \in \mathbb{R}^{m\times m}$ are the sample correlation matrices of $Z_s$ respectively. Similarly, let $D_X$ be the diagonal matrix containing the diagonal elements of $\mathbf{X}_b^\top \mathbf{X}_b$, then $\widehat{M}_X := D_X^{-1/2} \mathbf{X}_b^\top \mathbf{X}_b D_X^{-1/2} \in \mathbb{R}^{d\times d}$ is the sample correlation matrix of $X_b$. ♣

Due to the close relation of the joint summary statistics with the reduced form model (4.2.2) it is easier to develop methods for the joint summary statistics. However, in

most publicly available data (e.g., UK Biobank [2024] and GWAS Catalog [2024]) only the marginal summary statistics are available. Fortunately, it is possible to transform marginal summary statistics into joint summary statistics. This means that any theoretical developments that apply to one also apply to the other. The exact correspondence is given in the following proposition.

**Proposition 4.2.3** (Marginal to joint summary statistics)**.** *Assume we are given* $\mathcal{D}_{a,b}^{marginal} = \{\widehat{\eta}, \widehat{\sigma}_\eta^2, \widehat{H}, \widehat{\sigma}_H^2, \widehat{M}_{Z_a}, \widehat{M}_{Z_b}, \widehat{M}_X\}$. *Define diagonal matrices* $D_a, D_b^{(1)}, \ldots, D_b^{(m)} \in \mathbb{R}^{m\times m}$ *such that for all* $k, i \in [m]$, $(D_a)_i^i := (\widehat{\sigma}_{\eta,i}^2 + (\widehat{\eta}_i)^2)^{1/2}$ *and* $(D_b^{(k)})_i^i := ((\widehat{\sigma}_{H,i}^k)^2 + (\widehat{H}_i^k)^2)^{1/2}$. *Then it holds for all* $k, l \in [d]$ *that*

- $\widehat{\pi} = D_a(D_a\ \widehat{M}_{Z_a})^{-1}\ \widehat{\eta}$,
- $\widehat{\Sigma}_\pi = (1 - \widehat{\eta}^\top D_a \widehat{M}_{Z_a}^{-1} D_a \widehat{\eta}) D_a \widehat{M}_{Z_a}^{-1} D_a$,
- $\widehat{\Pi}^k = D_b^{(k)}(D_b^{(k)}\ \widehat{M}_{Z_b})^{-1}\ \widehat{H}^k$, *and*
- $\widehat{\Sigma}_\Pi^{[kl]} = (\widehat{M}_{X,k}^l - \widehat{H}^{k\top} D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)} \widehat{H}^l) D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)}$.

In practice, one often does not observe both $\widehat{M}_{Z_a}$ and $\widehat{M}_{Z_b}$ and instead only observes a single estimate that converges to the correlation of $Z$. In such cases, it can be shown that using the same transformation as in Proposition 4.2.3 is asymptotically equivalent to working with the joint summary statistics.

### 4.2.1 Identifiability via sparsity under the reduced IV model

For the causal effect $\beta^*$ to be identified, the number of instruments is usually required to be no less than the number of covariates. In the one-sample individual-level data setting, this can be seen from the solution space based on the IV moment condition under the SCM (4.2.1),

$$\mathcal{B}^{\text{ind}} = \left\{\beta \in \mathbb{R}^d : \mathbb{E}(Z\,Y) = \mathbb{E}(ZX^\top)\beta)\right\}. \tag{4.2.3}$$

This space is in general non-degenerate if the dimension of the instruments is larger than the number of covariates. When the causal effect is sparse, however, it is possible to allow more covariates than instruments.

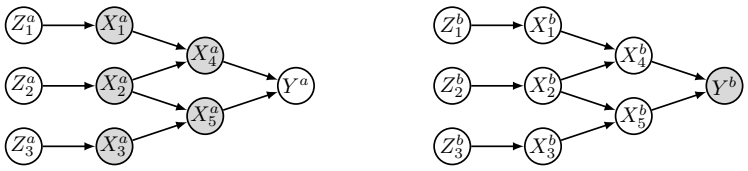



Figure 4.2.1: An example of two-sample IV scenario that is considered as underidentified in the usual sense. Hidden confounders between $X$ and $Y$ are omitted for clarity. While the two DAGs have the same structure, in sample $a$ (left) the covariates $X$ are not observed and in sample $b$ (right) the outcome $Y$ is not observed, these unobserved variables are represented by gray nodes.

Pfister and Peters [2022] study in detail the identifiability conditions under the SCM (4.2.1). In the following, we describe the identifiability conditions under the reduced model (4.2.2) which is compatible with the two-sample summary statistics scenario. Lemma 4.2.4 describes the solution space of the causal effects with the reduced form model.

**Lemma 4.2.4.** *If $\mathbb{E}[ZZ^\top]$ has full rank, the solution space of the causal effects based on the IV moment condition can be written as*

$$\mathcal{B}^{sum} = \left\{\beta \in \mathbb{R}^d : \pi - \Pi\beta = 0\right\}. \tag{4.2.4}$$

We will focus on the case where the instruments do not have direct effects on the response, which is implied by the SCMs (4.2.1) and its reduced form (4.2.2). This is usually referred to as the exclusion restriction criteria of IV. In genetics research, such a direct effect is also referred to as pleiotropy [see e.g., Hemani et al., 2018]. We demonstrate empirically with additional simulations in Supplementary Material 4.E that the proposed methods still perform well under small violations of this assumption. An example of the possible scenario represented by directed acyclic graphs (DAGs) is given in Figure 4.2.1. As we will see shortly, although the number of instrument is less than the number of covariates in this case, the causal effect from $X$ to $Y$ may still be identified.

Under the two-sample summary statistic setting, the identifiability conditions in Pfister and Peters [2022] can be written as follows.[2]

**Assumption 1.** *For all $S \subseteq [d]$, let $\Pi^S$ be the submatrix of $\Pi$ containing the columns in $S$. We assume the following regarding the true parameter $\Pi$*

*(a)* *rank$(\Pi^{PA(Y)}) = |PA(Y)|$.*

*(b)* *$\forall S \subseteq [d]$, it holds that rank$(\Pi^S) \leq$ rank$(\Pi^{PA(Y)})$ and Im$(\Pi^S) \neq$ Im$(\Pi^{PA(Y)})$ imply that $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{PA(Y)}(\beta^*)^{PA(Y)}$.*

*(c)* *$\forall S \subseteq [d]$ with $|S| = |PA(Y)|$ and $S \neq PA(Y)$, Im$(\Pi^S) \neq$ Im$(\Pi^{PA(Y)})$.*

To obtain a sparse solution, it is natural to consider the following optimization problem

$$\min_{\beta \in \mathcal{B}^{\text{sum}}} ||\beta||_0. \tag{4.2.5}$$

Theorem 4.2.5 shows that under Assumption 1, $\beta^*$ is a unique solution to (4.2.5). The proof follows similarly as in Pfister and Peters [2022, Theorem 3].

**Theorem 4.2.5** (Identifiability of sparse causal effect with reduced form model)**.** *If Assumption 1 (a) and (b) hold, then $\beta^*$ is a solution to (4.2.5). If in addition Assumption 1 (c) holds, then $\beta^*$ is the unique solution.*

[2] For a matrix $A$, Im$(A)$ denotes the image of $A$.

### 4.2.2 Anderson-Rubin test for two-sample summary statistics

The AR test is a well-known weak-instrument robust test for the causal effect, and the LIML estimator is known to minimize the AR statistic [e.g., Dhrymes, 2012]. Wang and Kang [2022] consider the two-sample summary statistic version of the AR test when there is a single covariate, which can be seen as a a generalization of the the modified Q statistic proposed by Bowden et al. [2019] for independent instruments. The following result is a generalization of Wang and Kang [2022] in the presence of multiple covariates[3], which will be referred to as the Q statistic.

**Theorem 4.2.6** (Q statistic)**.** *Assume Assumption 4.5.1 holds. For all $\beta \in \mathbb{R}^d$, define the Q statistic as*

$$Q(\beta) := (\widehat{\pi} - \widehat{\Pi}\beta)^\top (\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta))^{-1}(\widehat{\pi} - \widehat{\Pi}\beta), \tag{4.2.6}$$

*where $\widehat{\Sigma}_\Pi(\beta) := \xi(\beta)\widehat{\Sigma}_\Pi\xi^\top(\beta)$ with $\xi(\beta) := \beta^\top \otimes I_m$. Then it holds for all $\beta \in \mathbb{R}^d$ and all $r \in (0, \infty)$ that*

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{t \in \mathbb{R}} \sup_{\substack{P \in \mathcal{P}: \\ \beta \in \mathcal{B}^{sum}(P)}} |\mathbb{P}_P\left(Q(\beta) \leq t\right) - \kappa_m(t)| = 0,$$

*where $\kappa_m$ is the CDF of the chi-squared distribution with $m$-degrees of freedom.*

The Q statistic is the two-sample counterpart of the one-sample AR statistic, we present their connections in Supplementary Material 4.A.1. Its minimizer can also be viewed as a generalized method of moments (GMM) estimator [Hansen, 1982], and it is related to the J statistic in economics literature. See Remark 4.5.1 in the Supplementary Material for additional comments on the definition of $\widehat{\Sigma}_\Pi(\beta)$.

## 4.3 Estimating sparse causal effects with spaceTSIV

We describe two estimation procedures to the optimization problem (4.2.5). The first procedure is the two-sample summary statistics counterpart of `spaceIV` by Pfister and Peters [2022], and the second procedure employs L1-penalization to replace subset selection which has the advantage of faster computational speed. For both procedures, we will use the following estimator, which is the minimizer of the Q statistic constrained on a specific support. For all $S \subseteq \{1, \dots, d\}$, define

$$\widehat{\beta}^Q(S) := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d:\ \mathrm{supp}(\beta) = S} Q(\beta). \tag{4.3.7}$$

In order to provide precise theoretical results, we further let $\mathcal{P}$ denote a family of distributions for $(X, Y, Z)$ generated by (4.2.1) which is assumed to be sufficiently regular (see Assumption 4.5.1 for details). For all $P \in \mathcal{P}$, we let $\beta^*(P)$ denote the causal effect and $\mathcal{B}^{\mathrm{sum}}(P)$ be the subset $\mathcal{B}^{\mathrm{sum}}$ induced by the distribution $P$ (both of which are fully identified from the observational distribution $P$).

[3] A related result is also considered by Patel et al. [2024] where a dispersion parameter is included and the principal components of the instruments are used. Here we focus on the case where the instruments are valid.

### 4.3.1 Sparsity by subset selection

For all $s \in [d]$, let

$$\widehat{\beta}^Q(s) \coloneqq \widehat{\beta}^Q\Big(\arg\min_{S\subseteq\{1,\dots,d\}:|S|=s} Q\left(\widehat{\beta}^Q(S)\right)\Big).$$

Moreover, following Theorem 4.2.6, for all $s \in [d]$ and for all $\alpha \in (0,1)$, the hypothesis test

$$\phi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}}) \coloneqq \mathbb{1}\left(Q(\widehat{\beta}^Q(s)) > \kappa_m^{-1}(1-\alpha)\right)$$

has uniform asymptotic level for the null hypothesis

$$\mathcal{H}_0(s) \coloneqq \{P \in \mathcal{P} \mid \exists \beta \in \mathcal{B}^{\text{sum}}(P) : \|\beta\|_0 = s\},$$

that is, for $\alpha \in (0,1)$, it holds that

$$\lim_{n_a,n_b\to\infty} \sup_{P\in\mathcal{H}_0(s)} \mathbb{P}_P(\phi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}}) = 1) \le \alpha.$$

An algorithm defining the `spaceTSIV` estimator using subset selection is given in Algorithm 6. Theorem 4.3.1 shows that it is consistent.

**Theorem 4.3.1.** *Assume Assumption 4.5.1 holds. Let $\mathcal{D}_{a,b}^{joint}$ be the joint summary statistics based on two independent samples of size $n_a$ and $n_b$ respectively. Let $P \in \mathcal{P}$ and $s_{max} \in \mathbb{N}$ such that $s_{max} \ge \|\beta^*(P)\|_0$. If Assumption 1 (a) and (b) holds, then for all $r \in (0,\infty)$*

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \mathbb{P}_P\left(\|\widehat{\beta}^{\le s_{max}}\|_0 = \|\beta^*\|_0\right) \ge 1-\alpha;$$

*if in addition Assumption 1 (c) also holds, then for all $\varepsilon > 0$ and all $r \in (0,\infty)$*

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \mathbb{P}_P\left(\|\widehat{\beta}^{\le s_{max}} - \beta^*\|_2 < \varepsilon\right) \ge 1-\alpha.$$

### 4.3.2 Sparsity by L1 penalty

The subset selection approach introduced in Section 4.3.1 becomes computationally infeasible when the number of covariates is large. We therefore propose a faster approach that uses L1 penalization to estimate the support of $\beta^*$ and then adapt the testing procedure from the previous section. More specifically, for a penalty parameter $\lambda > 0$, we first minimize the following L1-loss

$$\mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta) = \frac{1}{2}\|\widehat{\pi} - \widehat{\Pi}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{4.3.8}$$

Define $\widehat{\beta}(\lambda) \coloneqq \arg\min_{\beta\in\mathbb{R}^d} \mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)$ and $\widehat{S}_\lambda \coloneqq \text{supp}(\widehat{\beta}(\lambda))$. We then propose to refit the parameter as in (4.3.7) using the set $\widehat{S}_\lambda$ and performing the hypothesis test defined by

$$\phi_\lambda^\alpha(\mathcal{D}_{n_a,n_b}^{\text{joint}}) \coloneqq \mathbb{1}\left(Q(\widehat{\beta}^Q(\widehat{S}_\lambda)) > \kappa_m^{-1}(1-\alpha)\right).$$

**Algorithm 6:** `spaceTSIV` with L0 penalization

**Input:** Joint summary statistics $\mathcal{D}_{a,b}^{\text{joint}}$, maximum support size $s_{\max}$, significance level $\alpha \in (0,1)$

Initialize $s \leftarrow 1$ and $\varphi \leftarrow 1$
**while** $s \leq s_{max}$ *and* $\varphi = 1$ **do**
  Set $\mathbf{S}_s$ to the set of all subsets of $[d]$ of size $s$
  **for** $S \in \mathbf{S}_s$ **do**
    Compute $\widehat{\beta}^Q(S)$
    Compute $Q(\widehat{\beta}^Q(S))$
  $S_{\text{best}} \leftarrow \arg\min_{S \in \mathbf{S}_s} Q(\widehat{\beta}^Q(S))$
  $\widehat{\beta}(s) \leftarrow \widehat{\beta}^Q(S_{\text{best}})$
  $\varphi \leftarrow \phi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}})$
  $s \leftarrow s + 1$
$\widehat{\beta}_{\leq s_{\max}} \leftarrow \widehat{\beta}(s)$

**Output:** Final estimate $\widehat{\beta}_{\leq s_{\max}}$ and test result $\varphi$

By similar arguments as in Section 4.3.1 this test for $S = \widehat{S}_\lambda$ has uniform asymptotic level for the null hypothesis

$$H_0(S) := \{P \in \mathcal{P} \mid \exists \beta \in \mathcal{B}^{\text{sum}}(P) : \text{supp}(\beta) = S\}.$$

Under sufficient regularity conditions and assuming that $\beta^*$ is indeed sparse, one can hope—based on similar results for high-dimensional linear models [e.g., Bühlmann and Van De Geer, 2011]—that for appropriately chosen $\lambda$ it holds that $\widehat{S}_\lambda$ converges to $\text{supp}(\beta^*)$. This motivates the following estimator, `spaceTSIV` with L1 penalization, defined in Algorithm 7.

**Algorithm 7:** `spaceTSIV` with L1 penalization

**Input:** Joint summary statistics $\mathcal{D}_{a,b}^{\text{joint}}$, a vector of penalty values in decreasing order $\{\lambda_1, \ldots, \lambda_\ell\}$, significance level $\alpha \in (0,1)$

Initialize $l \leftarrow 1$ and $\varphi \leftarrow 1$
**while** $l \leq \ell$ *and* $\varphi = 1$ **do**
  $\lambda \leftarrow \lambda_l$
  $\widehat{S}_\lambda \leftarrow \text{supp}\left(\arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)\right)$
  Compute $\widehat{\beta}^Q(\widehat{S}_\lambda)$
  $\varphi \leftarrow \phi_\lambda^\alpha(\mathcal{D}_{a,b}^{\text{joint}})$
  $l \leftarrow l + 1$
$\widehat{\beta}_{\leq \lambda_{\max}} \leftarrow \widehat{\beta}(\widehat{S}_\lambda)$

**Output:** Final estimate $\widehat{\beta}_{\leq \lambda_{\max}}$ and test result $\varphi$

Intuitively, if the subset selection is indeed correct (i.e., it recovers the support of $\beta^*$) for the first accepted set, then this procedure should correctly estimate $\beta^*$. A full theoretical analysis, however, goes beyond the scope of this work and we propose this procedure only as a heuristic computational speed up.

### 4.3.3 Practical considerations

When using the subset selection approach in practice, it can happen that there are multiple estimates with different support of the same (smallest) size not being rejected by $\phi_s^\alpha$. This indicates, that at least in finite sample, the causal effect $\beta^*$ is not fully identified. We recommend reporting all subsets of the smallest size that are not rejected by $\phi_s^\alpha$ as possible effects.

Moreover, since the estimator `spaceTSIV` is based on optimizing a test statistics, one immediate approach to construct confidence intervals (CIs) is by inverting the test. In the real application, we construct the CIs for the non-zero causal effects by inverting $\phi_s^\alpha$ or $\phi_\lambda^\alpha$ and projecting onto each non-zero coordinate. We choose this approach for its practicality, but other approaches exist which may be more suitable [e.g., Londschien and Bühlmann, 2024], and one should also take into account the effect of post-selection inference [e.g., Lee et al., 2016]. One of the advantages of inverting the test is that is takes into account the strength of the instruments (and hence identifiability). So if the resulting CIs are unbounded this generally indicates that there is limited identifiability. This is well known property for the AR test [e.g., Dufour, 1997, Davidson and MacKinnon, 2014].

## 4.4 Experiments

Code for reproducing the simulations and the real-data application along with the data are available in the GitHub repository `https://github.com/shimenghuang/spacetsiv`. All experiments were run on a MacBook Pro laptop with M1 chip.

### 4.4.1 Simulations

We present simulation results for two data generating processes (DGPs) summarized below in this section. Further simulation results are provided in Supplementary Material 4.E. The first, DGP1, is a low dimensional example taken from Pfister and Peters [2022, Figure 3]. We compare the subset selection and the L1-penalization versions of `spaceTSIV`, denoted as `spaceTSIV-L0` and `spaceTSIV-L1` respectively, as well as the `TSIV` estimator (defined as the minimizer of (4.3.8) with $\lambda = 0$, in which case the generalized inverse is used). The second, DGP2, illustrates the scenario with higher dimensional covariates, sparser causal effects, and correlated instruments. In this setting we omit `spaceTSIV-L0` from the comparison due its high computational cost. Overview of the simulation setup is given below and more details can be found in Supplementary Material 4.E.1.

**DGP1 overview:** $m = 3$ and $d = 5$ and $\|\beta^*\|_0 = 2$. For increasing $n = n_a = n_b$, we generate iid $\{(Y_i, Z_i)\}_{i=1}^{n_a}$ and $\{(X_i, Z_i)\}_{i=1}^{n_b}$ according to a linear SCM with Gaussian errors and then compute the summary statistics using seemingly unrelated regression.
**DGP2 overview:** $m = 5$, $d = 100$, and $\|\beta^*\|_0 = 2$. With fixed values of $\pi$, $\Pi$, $\Sigma_\pi$, and $\Sigma_\Pi$, and increasing $n = n_a = n_b$, we generate $\widehat{\pi}_{n_a} \sim \mathcal{N}(\pi, \frac{1}{n_a}\Sigma_\pi)$ and $\widehat{\Pi}_{n_b} \sim \mathcal{N}(\Pi, \frac{1}{n_b}\Sigma_\Pi)$, and set $\widehat{\Sigma}_{\pi,n_a} = \Sigma_\pi$ and $\widehat{\Sigma}_{\Pi,n_b} = \Sigma_\Pi$.

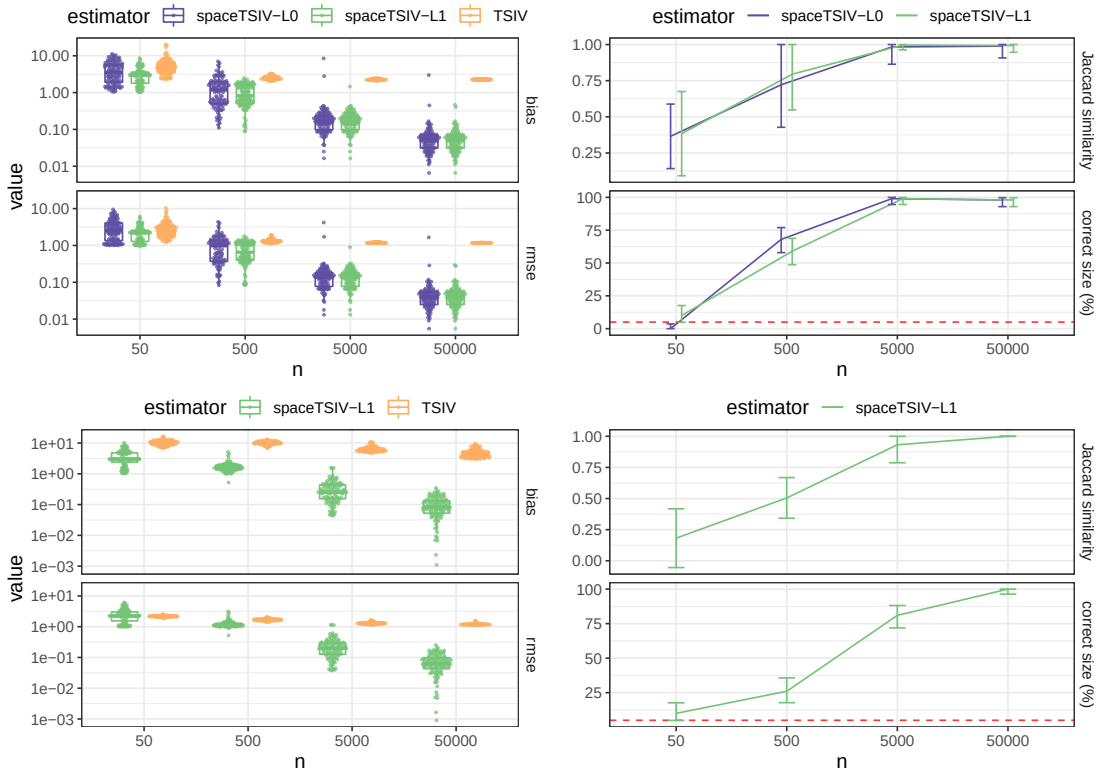



Figure 4.4.2: Results using data generated by DGP1 (top) and DGP2 (bottom) based on 100 repetitions. Left: Bias and rmse of the estimators. The y-axis is on log scale for clarity. Right: Average Jaccard similarity between the selected covariates and the true causal covariates (error bars indicate confidence intervals constructed by mean plus/minus one standard error), and percentage of estimates that have the correct support size (error bars indicate 95% binomial confidence intervals).

We evaluate `spaceTSIV` based on both its variable selection and estimation performances. The results are shown in Figure 4.4.2. We can see that the bias and rmse of `spaceTSIV` shrinks with increasing sample size with either L0 or L1 penalization, which is not the case for the non-sparse estimator `TSIV`. In terms of variable selection, we see

that for both DGPs as the sample sizes increase, the Jaccard similarity[4] increases to around 1, and the percentage of estimates having the correct support size also increases to around 100%, empirically confirming the consistency results in Theorem 4.3.1. The performance of `spaceTSIV-L0` and `spaceTSIV-L1` are similar in terms of both estimation and variable selection for DGP1.

### 4.4.2 Application

We apply our methods to summary statistics of SNP-level associations where the covariates and the outcome come from two separate GWAS sources. The covariates' summary statistics come from the GTEX consortium, which measure levels of expression of protein coding genes across multiple tissue types in the human body. Gene expression is a convenient and reliable upstream marker of protein production, which would be the natural target of a future drug. We specifically focus on expression of the GLP1R gene in 10 tissue types that are relevant to the treatment of cardio-metabolic disease. These are brain caudate, hypothalamus, atrial appendage, left ventricle, lung, nerve, pancreas, stomach, testis, and thyroid. The SNP-outcome summary statistics measure the genetic association with coronary artery disease (CAD) risk, and are obtained from the CARDIoGRAMplusC4D consortium. These data were first analysed in in Patel et al. [2024], who proposed a novel principle component analysis (PCA) method for constructing orthogonal composite instruments from 851 SNPs in the GLP1R gene region. For this analysis, they use 23 principle components (PCs) as IVs for the 10 covariates. The analysis by Patel et al. [2024] suggests that GLP1R expression only has a significant effect on CAD risk in 2 of the 10 tissues, although this was based on 95% confidence intervals using a normal approximation which, unlike the test-inversion method we use, does not always reliably capture the true uncertainty of IV estimates when the instruments are weak.

Based on the analysis of Patel et al. [2024], it is reasonable to believe that the causal effects are sparse in this application. Rather than opting for PCA pre-processing of the genetic summary statistics, we consider the selection of individual SNPs instruments based on the more conventional approach using the first-stage F-statistics[5] of the gene expression summary statistics. We keep the top two genetic variants with the largest first-stage F-statistics for each of the 10 covariates. Since some SNPs are most strongly associated with multiple covariates, we eventually keep 17 of the 851 genetic variants in the original data. Moreover, since the summary statistic data contains only the marginal associations along with their standard errors, we use the adjustment method in Proposition 4.2.3 to obtain the estimated joint effects and variance-covariance matrices.

The analysis results based on `spaceTSIV` with L0 and L1 penalization and regular `TSIV` are reported in Figure 4.4.3, where the 90% CIs are obtained from inverting $\phi_s^\alpha$ and $\phi_\lambda^\alpha$ as described in Section 4.3.3. They show that the CIs for `TSIV` are all of infinite length. This

[4] For two sets $A$ and $B$, the Jaccard similarity is defined as $\text{Jaccard}(A, B) \coloneqq \dfrac{|A \cap B|}{|A \cup B|}$.

[5] Given a marginal OLS coefficient $\widehat{\gamma} \in \mathbb{R}$ and its corresponding standard error $\widehat{\sigma} \in \mathbb{R}$, the first-stage F-statistic is defined as $\widehat{\gamma}^2/\widehat{\sigma}^2$.

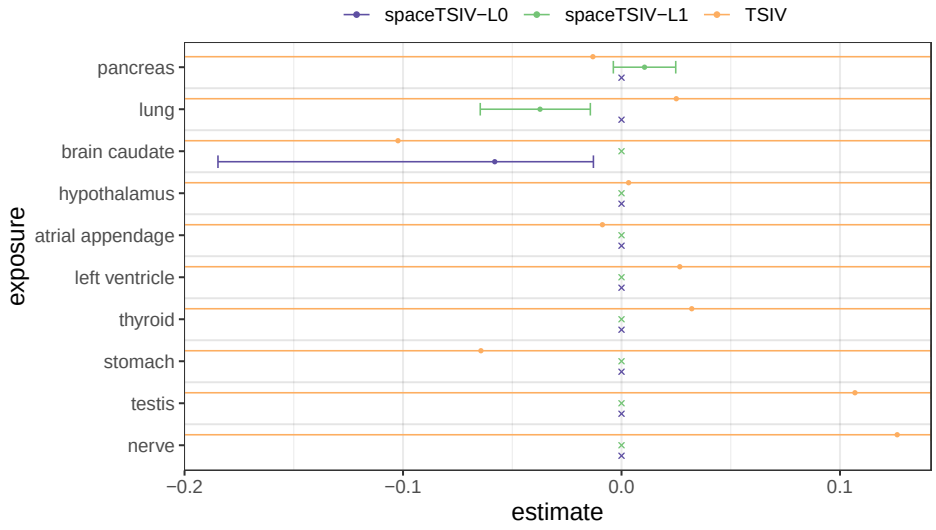



Figure 4.4.3: Estimated effects of the GLP1R expression in 10 tissues using the selected 17 genetic variants as instruments. Error bars represent 90% confidence intervals (CIs) constructed by inverting $\phi_s^\alpha$ and $\phi_\lambda^\alpha$ respectively, and projecting onto each coordinate.

demonstrates that, even though there are more instruments than covariates, the causal effects are still under-identified due to weak instruments. Moreover, the `spaceTSIV` with L0 penalization yields a single set of size 1 while with L1 penalization we obtain a set of size 2. The significant negative effect of brain caudate aligns with the analysis result in Patel et al. [2024] and is biologically meaningful. The different result from `spaceTSIV-L1` could be due to the high correlation of the SNPs, which may result in the L1 relaxation of the L0 minimization problem not achieving the same estimate. In general we recommend using the L0 procedure whenever computationally feasible as it comes with clear theoretical guarantees.

## 4.5 Discussion

We propose `spaceTSIV` for sparse multivariable causal effect estimation under unobserved confounding, which is applicable to the two-sample summary statistics setting. Two methods using subset selection and L1-penalization respectively are provided. We prove consistency for the subset selection approach and illustrate the results in simulations. We also show in simulations that the L1-penalization approach, which is much more computational efficient, can achieve similar performance as the subset selection approach in terms of bias and consistency. To focus on the main idea of this work, we have assumed that the summary statistics utilized in the analysis are obtained from two independent and homogeneous samples, which is commonly assumed in genetic epidemiology. However, it would be interesting to generalize the methods to heterogeneous

samples similar to results by Zhao et al. [2019] in the non-sparse setting. Moreover, if the summary statistics are obtained from two samples with overlapping observations, additional correlations should be taken into account.

## Acknowledgement

The authors would like to thank Stephen Burgess and Ashish Patel for helpful discussions at the start of this research project, and Anton Rask Lundborg for helpful discussions on the uniform asymptotic results. SH and NP are supported by a research grant (0069071) from Novo Nordisk Fonden. JB is funded at the University of Exeter by research grant MR/X011372/1.

# Supplementary material for "Sparse causal effect estimation using two-sample summary statistics in the presence of unmeasured confounding"

## 4.A. Details of test statistics and test-based estimators

### 4.A.1 Connection between Anderson-Rubin statistic and Q statistic

Suppose we observe one set of iid samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$. The AR statistic is given by

$$AR(\beta) \coloneqq \frac{n-m}{m} \cdot \frac{(\mathbf{Y} - \mathbf{X}\beta)^\top P_Z (\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)}, \tag{4.A.1}$$

where $P_Z = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ and $M_Z = I_d - P_Z$. When the true causal effect $\beta^*$ is identified, $m \cdot AR(\beta^*) \xrightarrow{d} \chi^2_m$ [Anderson and Rubin, 1949, Londschien and Bühlmann, 2024].

We can rewrite the AR statistic in terms of (joint) OLS estimates and their respective estimated variance-covariance matrices. Specifically, let $\widehat{\pi} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}\mathbf{Y}$ and $\widehat{\Pi} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}\mathbf{X}$, we have that

$$(\mathbf{Y} - \mathbf{X}\beta)^\top P_Z (\mathbf{Y} - \mathbf{X}\beta) = (\widehat{\pi} - \widehat{\Pi}\beta)^\top (\mathbf{Z}^\top\mathbf{Z})(\widehat{\pi} - \widehat{\Pi}\beta).$$

Moreover, for all $\beta \in \mathbb{R}^d$ define

$$\begin{aligned}
\widehat{\Sigma}_\pi &= (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi})(\mathbf{Z}^\top\mathbf{Z})^{-1}, \\
\widehat{\Sigma}_\Pi(\beta) &= \beta^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})\beta(\mathbf{Z}^\top\mathbf{Z})^{-1}, \text{ and} \\
\widehat{\Sigma}_{\pi,\Pi}(\beta) &= (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})\beta(\mathbf{Z}^\top\mathbf{Z})^{-1}.
\end{aligned}$$

Then, we can expand the denominator in (4.A.1) as follows

$$\begin{aligned}
&(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta) \\
&\quad = (M_Z\mathbf{Y} - M_Z\mathbf{X}\beta)^\top (M_Z\mathbf{Y} - M_Z\mathbf{X}\beta) \\
&\quad = (\mathbf{Y} - \mathbf{Z}\widehat{\pi} - (\mathbf{X} - \mathbf{X}\widehat{\Pi}\beta))^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi} - (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta) \\
&\quad = (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi}) + \beta^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta - 2(\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta,
\end{aligned}$$

which implies

$$(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Z}^\top\mathbf{Z})^{-1} = \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta) - 2\widehat{\Sigma}_{\pi,\Pi}(\beta).$$

Therefore,

$$\begin{aligned}
AR(\beta) &= \frac{(\widehat{\pi} - \widehat{\Pi}\beta)^\top (\mathbf{Z}^\top\mathbf{Z})(\widehat{\pi} - \widehat{\Pi}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)} \\
&= \frac{1}{m}(\widehat{\pi} - \widehat{\Pi}\beta)^\top \left( \frac{1}{n-m}\widehat{\Sigma}_\pi + \frac{1}{n-m}\widehat{\Sigma}_\Pi(\beta) - \frac{2}{n-m}\widehat{\Sigma}_{\pi,\Pi}(\beta) \right)^{-1} (\widehat{\pi} - \widehat{\Pi}\beta).
\end{aligned}$$

From this expression, we can see the connection between the AR statistic and the Q statistic. Specifically, for a fixed $m$ and large $n$, the difference between $mAR(\beta)$ and $Q(\beta)$ is the term $\widehat{\Sigma}_{\pi,\Pi}$, which is related to the covariance between the residuals of a $Y$ on $Z$ and a $X$ on $Z$ regression. In the two sample setting this covariance is zero because the two regressions are performed on independent samples and therefore it is not needed in the the Q statistic.

### 4.A.2 Coordinate descent for minimizing the TSIV-L1 loss

We describe the coordinate descent procedure for minimizing (4.3.8). Let

$$\mathcal{L}^{\text{TSIV}}(\beta) = \frac{1}{2}\|\widehat{\pi} - \widehat{\Pi}\beta\|_2^2.$$

For a matrix $A$, denote $A^{-j}$ as the matrix removing A's $j$-th column. The derivative of $\mathcal{L}^{\text{TSIV}}(\beta)$ w.r.t $\beta_j$ is

$$\begin{aligned}\frac{\partial \mathcal{L}^{\text{TSIV}}(\beta)}{\beta_j} &= -(\widehat{\Pi}^j)^\top(\widehat{\pi} - \widehat{\Pi}\beta) \\ &= -(\widehat{\Pi}^j)^\top\widehat{\pi} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^{-j}\beta_{-j} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^j\beta_j \\ &= -\rho_j + \eta_j\beta_j\end{aligned} \tag{4.A.2}$$

where $\rho_j \coloneqq (\widehat{\Pi}^j)^\top\widehat{\pi} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^{-j}\beta_{-j}$ and $\eta_j \coloneqq (\widehat{\Pi}^j)^\top\widehat{\Pi}^j$. The subgradient of $\lambda||\beta||_1$ w.r.t $\beta_j$ is

$$\frac{\partial \lambda||\beta||_1}{\beta_j} = \frac{\partial \lambda|\beta_j|}{\beta_j}\begin{cases}\{-\lambda\} & \beta_j < 0 \\ [-\lambda, \lambda] & \beta_j = 0 \\ \{-\lambda\} & \beta_j > 0\end{cases} \tag{4.A.3}$$

Combining (4.A.2) and (4.A.3), we have that the subgradient of $\mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)$ w.r.t. $\beta_j$ is

$$\frac{\partial \mathcal{L}_\lambda^{\text{TSIV}}(\beta)}{\partial \beta_j} = \begin{cases}-\rho^j + \eta^j\beta_j - \lambda & \beta_j < 0 \\ [-\rho^j - \lambda, -\rho^j + \lambda] & \beta_j = 0 \\ -\rho^j + \eta^j\beta_j + \lambda & \beta_j > 0.\end{cases} \tag{4.A.4}$$

Starting from an initial value of $\widehat{\beta}$, we loop through $j \in [J]$ and update the value of $\widehat{\beta}_j$ by solving the equation resulting from setting (4.A.4) to 0, which gives

$$\widehat{\beta}_j = \begin{cases}\frac{\rho^j + \lambda}{\eta^j} & \rho^j < -\lambda \\ 0 & -\lambda < \rho^j < \lambda \\ \frac{\rho^j - \lambda}{\eta^j} & \rho^j > \lambda.\end{cases}$$

## 4.B. Regularity conditions

**Assumption 4.5.1** (Regularity conditions)**.** *Let $\mathcal{P}$ be a family of distributions for $(X, Y, Z) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m$ generated by (4.2.1) and additionally satisfies that there exists $C_1, C_2, c, \eta > 0$ such that*

- $\sup_{P\in\mathcal{P}} \left(\mathbb{E}_P[\|X\|^{4+\eta}] + \mathbb{E}[\|Y\|^{4+\eta}] + \mathbb{E}_P[\|Z\|^{4+\eta}]\right) \leq c$
- $\inf_{P\in\mathcal{P}} \min\left(\lambda_{\min}(\mathbb{E}_P[ZZ^\top]), \lambda_{\min}(\mathbb{E}_P[XX^\top])\right) \geq C_1.$
- $\sup_{P\in\mathcal{P}} \max\left(\lambda_{\max}(\mathbb{E}_P[ZZ^\top]), \lambda_{\max}(\mathbb{E}_P[XX^\top])\right) \leq C_2.$

## 4.C. Proofs

### 4.C.1 Proof of Lemma 4.2.4

*Proof.* The following equivalences hold

$$\begin{aligned}
\beta \in \mathcal{B}^{\text{ind}} &\iff \mathbb{E}(ZY) = \mathbb{E}(ZX^\top)\beta \\
&\iff \mathbb{E}(ZZ^\top)\pi = \mathbb{E}(ZZ^\top)\Pi\beta \quad \text{since following (4.2.2), we have } \mathbb{E}(ZY) = \mathbb{E}(ZZ^\top)\pi \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and } \mathbb{E}(ZX^\top)\beta = \mathbb{E}(ZZ^\top)\Pi\beta \\
&\iff \pi = \Pi\beta \qquad \text{since } \mathbb{E}[ZZ^\top] \text{ is full rank.}
\end{aligned}$$

□

### 4.C.2 Proof of Proposition 4.2.3

*Proof.* We only prove the result for $\widehat{\Pi}$ and $\widehat{\Sigma}_\Pi$. $\widehat{\pi}$ and $\widehat{\Sigma}_\pi$ can be viewed as a special case of the former, with $d = 1$.

We first express $D_b^{(k)}$ in terms of the design matrices. To this end, observe that for all $j \in [m]$ and all $k \in [d]$, we have from the marginal OLS summary statistics that

$$\begin{aligned}
\left(\widehat{\sigma}^k_{\eta,j}\right)^2 &= \frac{(\mathbf{X}_b^k - \widehat{H}_j^k \mathbf{Z}_b^j)^\top (\mathbf{X}_b^k - \widehat{H}_j^k \mathbf{Z}_b^j)}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - 2\widehat{H}_j^k (\mathbf{X}_b^k)^\top \mathbf{Z}_b^j + (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - 2(\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j + (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} - (\widehat{H}_j^k)^2.
\end{aligned}$$

This further implies that $(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j = \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}{\left(\widehat{\sigma}_{\eta,j}^k\right)^2 + (\widehat{H}_j^k)^2}$, and hence $\left(\left(\widehat{\sigma}_{\eta,j}^k\right)^2 + (\widehat{H}_j^k)^2\right)^{-1} \widehat{H}_j^k = \frac{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k} \widehat{H}_j^k = \frac{(\mathbf{X}_b^k)^\top \mathbf{Z}_b^j}{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}$. Therefore, if we define the diagonal matrix $D_{Z_b}$ for all $i \in [m]$ by $(D_{Z_b})_i^i \coloneqq (\mathbf{Z}_b^i)^\top \mathbf{Z}_b^i$, it holds that

$$D_b^{(k)} = D_{Z_b}^{-1/2} \left((\mathbf{X}_b^k)^\top \mathbf{X}_b^k\right)^{1/2}.$$

Using this result, for all $k \in [d]$, we can expand the joint OLS estimate $\widehat{\Pi}^k$ as follows

$$\begin{aligned}
\widehat{\Pi}^k &= \left(\mathbf{Z}_b^\top \mathbf{Z}_b\right)^{-1} \mathbf{Z}_b^\top \mathbf{X}_b^k \\
&= \left(\mathbf{Z}_b^\top \mathbf{Z}_b\right)^{-1} D_{Z_b} \widehat{H}^k \\
&= D_{Z_b}^{-1/2} \widehat{M}_{Z_b}^{-1} D_{Z_b}^{-1/2} D_{Z_b} \widehat{H}^k \\
&= D_{Z_b}^{-1/2} \widehat{M}_{Z_b}^{-1} D_{Z_b}^{1/2} \widehat{H}^k \\
&= D_b^{(k)} \left(\widehat{D}_b^{(k)} M_{Z_b}\right)^{-1} \widehat{H}^k
\end{aligned}$$

Similarly, for all $k, l \in [d]$, the variance-covariance matrix between $\widehat{\Pi}^k$ and $\widehat{\Pi}^l$, can be expanded as follows.

$$\begin{aligned}
\widehat{\Sigma}_\Pi^{[kl]} &= \left(\mathbf{X}_b^k - \mathbf{Z}_b \widehat{\Pi}^k\right)^\top \left(\mathbf{X}_b^l - \mathbf{Z}_b \widehat{\Pi}^l\right) (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top \mathbf{X}_b^l - (\mathbf{X}_b^k)^\top \mathbf{Z}_b \widehat{\Pi}^l - (\widehat{\Pi}^k)^\top (\mathbf{Z}_b)^\top \mathbf{X}_b^l + (\widehat{\Pi}^k)^\top \mathbf{Z}_b^\top \mathbf{Z}_b \widehat{\Pi}^l\right) (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top \mathbf{X}_b^l - 2(\widehat{\Pi}^k)^\top \mathbf{Z}_b^\top \mathbf{Z}_b \widehat{\Pi}^l + (\widehat{\Pi}^k)^\top \mathbf{Z}_b^\top \mathbf{Z}_b \widehat{\Pi}^l\right) (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top \mathbf{X}_b^l - (\widehat{\Pi}^k)^\top \mathbf{Z}_b^\top \mathbf{Z}_b \widehat{\Pi}^l\right) (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1} \\
&= \left(\widehat{M}_{X,k}^l - \frac{\mathbf{Z}_b^\top \mathbf{X}_b^k}{(\mathbf{X}_k)^\top \mathbf{X}_b^k} \left(\frac{\mathbf{Z}_b^\top \mathbf{Z}_b}{\sqrt{(\mathbf{X}_k)^\top \mathbf{X}_b^k}\sqrt{(\mathbf{X}_b^l)^\top \mathbf{X}_b^l}}\right)^{-1} \frac{\mathbf{Z}_b^\top \mathbf{X}_b^k}{(\mathbf{X}_b^l)^\top \mathbf{X}_b^l}\right) \\
&\qquad \cdot \left(\frac{\mathbf{Z}_b^\top \mathbf{Z}_b}{\sqrt{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}\sqrt{(\mathbf{X}_b^l)^\top \mathbf{X}_b^l}}\right)^{-1} \\
&= \left(\widehat{M}_{X,k}^l - (\widehat{H}^k)^\top D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)} \widehat{H}^l\right) D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)}.
\end{aligned}$$

□

### 4.C.3 Proof of Theorem 4.2.5

*Proof.* *(First statement)* Assume Assumption 1 (a) and (b) hold. We would like to show that

$$\beta^* \in \arg\min_{\beta \in \mathcal{B}^{\text{sum}}} \|\beta\|_0.$$

Since $\beta^* \in \mathcal{B}^{\text{sum}}$, it suffices to show that for all $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, we have $\|\tilde{\beta}\|_0 \geq |\text{PA}(Y)|$. Fix a $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$. Since $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, it holds that $\pi = \Pi\tilde{\beta} = \Pi\beta^*$. Let $S \coloneqq \text{supp}(\tilde{\beta})$. Since $\forall j \in [d] \setminus \text{PA}(Y)$, $(\beta^*)^j = 0$, $\Pi\tilde{\beta} = \Pi\beta^*$ implies that

$$\Pi^S \tilde{\beta}^S = \Pi^{\text{PA}(Y)} (\beta^*)^{\text{PA}(Y)}. \tag{4.C.1}$$

For the sake of contradiction, suppose that $|S| < |\text{PA}(Y)|$. Then by Assumption 1 (a), we have that

$$\text{rank}(\Pi^{\text{PA}(Y)}) = \dim\left(\text{Im}(\Pi^{\text{PA}(Y)}\right) = |\text{PA}(Y)| > |S| \geq \dim\left(\text{Im}(\Pi^S)\right) = \text{rank}(\Pi^S).$$

This gives $\text{rank}(\Pi^{\text{PA}(Y)}) > \text{rank}(\Pi^S)$ which implies $\text{Im}(\Pi^{\text{PA}(Y)}) \neq \text{Im}(\Pi^S)$. Then by Assumption 1 (b), we have that $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}$, but this contradicts (4.C.1). This concludes the proof of the first statement.

*(Second statement)* It remains to show that there is no other solutions than $\beta^*$ when Assumption 1 (c) holds. Suppose for the sake of contradiction that there exists $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$ with $S \coloneqq \text{supp}(\tilde{\beta}) = |\text{PA}(Y)|$ and $S \neq \text{PA}(Y)$. Similarly as above, since $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, (4.C.1) holds. Then by Assumption 1 (c) we have $\text{Im}(\Pi^S) \neq \text{Im}(\text{PA}(Y))$. Moreover, by Assumption 1 (a) it holds that

$$\text{rank}(\Pi^{\text{PA}(Y)}) = |\text{PA}(Y)| = |S| \geq \text{rank}(\Pi^S).$$

Therefore, by Assumption 1 (b) $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}$, which again contradicts (4.C.1). This concludes the proof of the second statement. □

### 4.C.4 Proof of Theorem 4.2.6

*Proof.* First, observe that using $S_{n_a,n_b}$ as defined in Lemma 4.D.2, we can express the Q statistic for all $\beta \in \mathbb{R}^d$ as

$$Q(\beta) = S_{n_a,n_b}(\beta)^\top S_{n_a,n_b}(\beta).$$

Moreover, for all $\beta \in \mathcal{B}^{\text{sum}}$ it holds by definition that $\mu_{n_a,n_b} = 0$, hence Lemma 4.D.2 implies that $S_{n_a,n_b}(\beta)$ converges uniformly to a standard Gaussian distribution as $n_a, n_b$ tend to infinity and $n_a/n_b \to r$ for $r \in (0,\infty)$. Hence, by the continuous mapping theorem it holds that

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{\substack{P\in\mathcal{P}: \\ \beta\in\mathcal{B}^{\text{sum}}(P)}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(Q(\beta) \leq t) - \kappa_m(t)| = 0,$$

which completes the proof of Theorem 4.2.6. □

### 4.C.5 Proof of Theorem 4.3.1

*Proof.* Let $r \in (0, \infty)$ and assume that $n_a/n_b \to r$ throughout the proof. Using $S_{n_a,n_b}$ as defined in Lemma 4.D.2, we can express the Q statistic for all $\beta \in \mathbb{R}^d$ as

$$Q(\beta) = S_{n_a,n_b}(\beta)^\top S_{n_a,n_b}(\beta).$$

Furthermore, let $\overline{\mathcal{B}} \subseteq \mathbb{R}^d$ be a compact set and choose $\overline{\beta} \in \overline{\mathcal{B}}$ such that $\inf_{\beta\in\overline{\mathcal{B}}} \|S_{n_a,n_b}(\beta)\|_2^2 = \|S_{n_a,n_b}(\beta^*)\|_2^2$. Then, using standard probability bounds and dropping the $n_a, n_b$ from the notation for simplicity, we get for all $P \in \mathcal{P}$ and all $t \in [0, \infty)$ that

$$\begin{aligned}
&\mathbb{P}_P\left(\inf_{\beta\in\overline{\mathcal{B}}} \|S(\beta)\|_2^2 \le t\right) \qquad (4.C.2)\\
&= \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta}) + \mu(\overline{\beta})\|_2 \le \sqrt{t}\right)\\
&\le \mathbb{P}_P\left(\left|\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 - \|\mu(\overline{\beta})\|_2\right| \le \sqrt{t}\right)\\
&= \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 - \|\mu(\overline{\beta})\|_2 \le \sqrt{t}, \|S(\overline{\beta}) - \mu(\overline{\beta})\| \ge \|\mu(\overline{\beta})\|\right)\\
&\quad + \mathbb{P}_P\left(\|\mu(\overline{\beta})\|_2 - \|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \le \sqrt{t}, \|S(\overline{\beta}) - \mu(\overline{\beta})\| \le \|\mu(\overline{\beta})\|\right)\\
&\le \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\| \ge \|\mu(\overline{\beta})\|\right)\\
&\quad + \mathbb{P}_P\left(\|\mu(\overline{\beta})\|_2 - \|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \le \sqrt{t}\right)\\
&\le 2\mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \ge \|\mu(\overline{\beta})\|_2 - \sqrt{t}\right)\\
&\le 2\mathbb{P}_P\left(\sup_{\beta\in\overline{\mathcal{B}}} \|S(\beta) - \mu(\beta)\|_2 \ge \inf_{\beta\in\overline{\mathcal{B}}} \|\mu(\beta)\|_2 - \sqrt{t}\right). \qquad (4.C.3)
\end{aligned}$$

Next, observe that

$$\begin{aligned}
S(\beta) &= \sqrt{n_b}\left(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\right)^{-1/2}(\pi - \Pi\beta)\\
&= \sqrt{n_b}\left(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + (\beta/\|\beta\|_2)^\top\widehat{\Sigma}_X(\beta/\|\beta\|_2)\widehat{\Sigma}_{Z_b}^{-1}\right)^{-1/2}(\pi - \Pi(\beta/\|\beta\|_2)),
\end{aligned}$$

where $\widehat{\Sigma}_X \coloneqq \frac{1}{n_b}\sum_{i=1}^{n_b}(X_{bi} - \widehat{\Pi}^\top Z_{bi})(X_{bi} - \widehat{\Pi}^\top Z_{bi})^\top$. This in particular implies that $S$ and hence $Q$ does not depend on the norm of $\beta$. Moreover, for all $\beta \in \mathbb{R}^d$ with $\|\beta\|_2 = 1$

it holds that

$$\begin{aligned}
\|S(\beta)-\mu(\beta)\|_2 &= \sqrt{n_b}\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1})^{-1/2}((\pi-\Pi\beta)-(\widehat{\pi}-\widehat{\Pi}\beta))\|_2 \\
&\leq \sqrt{n_b}\|\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\|_{\mathrm{op}}^{-1/2}(\|\pi-\widehat{\pi}\|_2+\|\Pi\beta-\widehat{\Pi}\beta\|_2) \\
&\leq \left(\lambda_{\min}(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi)+\lambda_{\max}(\beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1})\right)^{-1/2}(\sqrt{n_b}\|\pi-\widehat{\pi}\|_2+\sqrt{n_b}\|\Pi-\widehat{\Pi}\|_{\mathrm{op}}) \\
&\leq \left(\lambda_{\min}(\widehat{\Sigma}_X)\lambda_{\max}(\widehat{\Sigma}_{Z_b}^{-1})\right)^{-1/2}(\sqrt{n_b}\|\pi-\widehat{\pi}\|_2+\sqrt{n_b}\|\Pi-\widehat{\Pi}\|_{\mathrm{op}}) \\
&\leq \left(\frac{\lambda_{\min}(\widehat{\Sigma}_{Z_b})}{\lambda_{\min}(\widehat{\Sigma}_X)}\right)^{1/2}(\sqrt{n_b}\|\pi-\widehat{\pi}\|_2+\sqrt{n_b}\|\Pi-\widehat{\Pi}\|_{\mathrm{op}}).
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and we used Weyl's inequality for the second inequality and that $\beta$ has norm one for the last inequality. Hence, using the bounds on the minimal eigenvalues of $\Sigma_X$ and $\Sigma_{Z_b}$ in Assumption 4.5.1, it holds that

$$\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1}\|S(\beta)-\mu(\beta)\|_2 = \mathcal{O}_{\mathcal{P}}(1) \tag{4.C.4}$$

as $n_a, n_b$ tend to infinity, where $\mathcal{O}_{\mathcal{P}}(1)$ denotes a uniformly bounded random variable with respect to $\mathcal{P}$. Finally, for all $s\in[d]$ define $\overline{\mathcal{B}}_s := \{\beta\in\mathbb{R}^d \mid \|\beta\|_0 = s \text{ and } \|\beta\|_2=1\}$. Then, using that $Q$ does not depend on the scale of $\beta$ and (4.C.3) we get that

$$\begin{aligned}
\mathbb{P}_P\left(\inf_{\beta:\|\beta\|_0=s}Q(\beta)\leq t\right) &= \mathbb{P}_P\left(\inf_{\beta\in\overline{\mathcal{B}}_s}Q(\beta)\leq t\right) \\
&\leq 2\mathbb{P}_P\left(\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1}\|S(\beta)-\mu(\beta)\|_2 \geq \inf_{\beta\in\overline{\mathcal{B}}_s}\|\mu(\beta)\|_2-\sqrt{t}\right).
\end{aligned} \tag{4.C.5}$$

Now for the first statement of Theorem 4.3.1, fix $s\in\mathbb{N}$ such that $s<\|\beta^*\|_0 = |\mathrm{PA}(Y)|$. It follows from Theorem 4.2.5 that for all $\beta\in\mathbb{R}^d$ with $\|\beta\|_0=\mathrm{s}$, $\pi-\Pi\beta\neq 0$. Therefore, there exists $\epsilon>0$ such that for all $\beta\in\mathbb{R}^d$ with $\|\beta\|_0=s$, it holds that $\|\pi-\Pi\beta\|_2>\epsilon$. Therefore, by (4.C.5) it holds that

$$\begin{aligned}
&\lim_{n_a,n_b\to\infty}\mathbb{P}_P\left(\phi_s(\mathcal{D}_{a,b}^{\mathrm{joint}})=1\right) \\
&= \lim_{n_a,n_b\to\infty}\mathbb{P}_P\left(\inf_{\beta:\|\beta\|_0=s}Q(\beta)>\kappa_m(1-\alpha)\right) \\
&\geq 1-\lim_{n_a,n_b\to\infty}2\mathbb{P}_P\left(\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1}\|S(\beta)-\mu(\beta)\|_2\geq\inf_{\beta\in\overline{\mathcal{B}}_s}\|\mu_{n_a,n_b}(\beta)\|_2-\sqrt{\kappa_m(1-\alpha)}\right) \\
&= 1,
\end{aligned}$$

where we used (4.C.4) together with

$$\begin{aligned}\lim_{n_a,n_b\to\infty}\inf_{\beta\in\overline{\mathcal{B}}_s}\|\mu_{n_a,n_b}(\beta)\|_2 &\geq \lim_{n_a,n_b\to\infty}\inf_{\beta\in\overline{\mathcal{B}}_s}\sqrt{n_b}\|\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi+\beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\|_{\mathrm{op}}^{-1/2}\epsilon\\ &\geq \lim_{n_a,n_b\to\infty}\inf_{\beta\in\overline{\mathcal{B}}_s}\left(\frac{1}{n_a}\lambda_{\max}(\widehat{\Sigma}_\pi)+\tfrac{1}{n_b}\frac{\lambda_{\max}(\widehat{\Sigma}_X)}{\lambda_{\min}(\widehat{\Sigma}_{Z_b})}\right)^{-1/2}\epsilon\\ &=\infty,\end{aligned}$$

where we again used the bounds on the minimal eigenvalues of $\Sigma_X$ and $\Sigma_{Z_b}$ in Assumption 4.5.1. Since this holds for all $s\in[d]$ with $s<\|\beta^*\|_0$, we further get

$$\begin{aligned}\lim_{n_a,n_b\to\infty}\mathbb{P}_P\left(\|\beta^{\leq s_{\max}}\|_0=\|\beta^*\|_0\right) &= \lim_{n_a,n_b\to\infty}\mathbb{P}_P\left(\min_{s<\|\beta^*\|_0}\phi_s=1 \text{ and } \phi_{\|\beta^*\|_0}=0\right)\\ &= \lim_{n_a,n_b\to\infty}\mathbb{P}_P\left(\phi_{\|\beta^*\|_0}=0\right)\\ &\geq 1-\alpha.\end{aligned}$$

For the second statement of Theorem 4.3.1, we can use the same argument. In this case, Theorem 4.2.5 implies that for all $c>0$ there exists $\epsilon>0$ such that for all $\beta\in\mathbb{R}^d$ with either $\|\beta\|_0<\|\beta^*\|_0$ or $\|\beta-\beta^*\|>\epsilon$ and $\|\beta\|_0=\|\beta^*\|_0$ it holds that $\|\pi-\Pi\beta\|_2>\epsilon$. Therefore, $\mu_{n_a,n_b}(\beta)$ again diverges and the arguments above remain valid. This completes the proof of Theorem 4.3.1. □

## 4.D. Additional results

*Remark* 4.5.1. In the definition of the (empirical) Q statistic in Theorem 4.2.6, we used

$$\widehat{\Sigma}_\Pi(\beta)\coloneqq\xi(\beta)\widehat{\Sigma}_\Pi\xi^\top(\beta) \tag{4.D.1}$$

where $\xi(\beta)\coloneqq\beta^\top\otimes I_m$. It follows from the properties of Kronecker product that (4.D.1) is equivalent to

$$\widehat{\Sigma}_\Pi(\beta)\coloneqq(\beta^\top(\mathbf{X}_b-\mathbf{Z}_b\widehat{\Pi})^\top(\mathbf{X}_b-\mathbf{Z}_b\widehat{\Pi})\beta)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1}, \tag{4.D.2}$$

which aligns with its population quantity $\Sigma_\Pi(\beta)\coloneqq(\beta^\top\mathbb{E}[u_b^X(u_b^X)^\top]\beta)\mathbb{E}[Z_bZ_b^\top]^{-1}$ used in Lemma 4.D.2, where $u_b^X$ is the population residual in (4.2.2). The reason why (4.D.1) is used instead of (4.D.2) in the Q statistic is that (4.D.1) only relies on the joint summary statistics, as the individual-level data is not available under the two-sample summary statistics setting. ♢

**Lemma 4.D.2.** *Assume Assumption 4.5.1. Let* $\mathcal{D}_{a,b}^{joint}=\left\{\widehat{\pi},\widehat{\Sigma}_\pi,\widehat{\Pi},\widehat{\Sigma}_\Pi\right\}$ *be the joint summary statistics based on two independent samples of sizes* $n_a$ *and* $n_b$*, respectively. For all* $\beta\in\mathbb{R}^d$*, define*

$$S_{n_a,n_b}(\beta)\coloneqq\left(\tfrac{1}{n_a}\widehat{\Sigma}_\pi+\tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta)\right)^{-1/2}(\widehat{\pi}-\widehat{\Pi}\beta)$$

*and*

$$\mu_{n_a,n_b}(\beta) \coloneqq \left(\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta)\right)^{-1/2}(\pi - \Pi\beta).$$

*Then, for all* $\beta \in \mathbb{R}^d$ *and all* $r \in (0,\infty)$ *it holds that*

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^m} \left|\mathbb{P}_P\left(S_{n_a,n_b}(\beta) - \mu_{n_a,n_b}(\beta) \le t\right) - \Phi_m(t)\right| = 0.$$

*Proof.* Fix an arbitrary $\beta \in \mathbb{R}^d$. Using by standard uniform convergence results for the OLS estimator [e.g., Lundborg et al., 2022, Lemma S10] it holds that

$$\sqrt{n_a}\Sigma_\pi^{-1/2}(\widehat{\pi} - \pi)$$

with $\Sigma_\pi \coloneqq \mathbb{E}[(u_a^Y)^2]\mathbb{E}[Z_a Z_a^\top]^{-1}$ (where $u_a^Y$ are the population residuals in (4.2.2) for sample $a$) converges uniformly w.r.t. $\mathcal{P}$ to a standard $m$-variate Gaussian distribution as $n_a$ tends to infinity. Similarly, when considering the regression of $\beta^\top X$ on $Z$, it holds that

$$\sqrt{n_b}\Sigma_\Pi(\beta)^{-1/2}(\widehat{\Pi} - \Pi)\beta$$

with $\Sigma_\Pi(\beta) \coloneqq (\beta^\top\mathbb{E}[u_b^X(u_b^X)^\top]\beta)\mathbb{E}[Z_b Z_b^\top]^{-1}$ (where $u_b^X$ are the residuals in (4.2.2) for sample $b$) converges uniformly w.r.t. $\mathcal{P}$ to a standard $m$-variate Gaussian distribution as $n_b$ tends to infinity. Combining these results and using that $n_a/n_b \to r$ and $\widehat{\pi}$ and $\widehat{\Pi}$ are estimated based on independent samples, we further have that

$$\sqrt{n_b}\left(\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta)\right)^{-1/2}\left((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)\right) \tag{4.D.3}$$

converges uniformly w.r.t. $\mathcal{P}$ to a standard $m$-variate Gaussian distribution as $n_a$ and $n_b$ tend to infinity.

Next, we show for all $\epsilon > 0$ that

$$\lim_{n_a\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\|\widehat{\Sigma}_\pi - \Sigma_\pi\|_{\mathrm{op}} > \epsilon\right) = 0 \text{ and } \lim_{n_b\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\|\widehat{\Sigma}_\Pi(\beta) - \Sigma_\Pi(\beta)\|_{\mathrm{op}} > \epsilon\right) = 0. \tag{4.D.4}$$

As the proofs for both results are the same we only show it for $\widehat{\Sigma}_\pi$. First, we express the estimator as

$$\widehat{\Sigma}_\pi = \frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai} - \widehat{\pi}^\top Z_{ai})^2\left(\frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top\right)^{-1}.$$

We now consider the two product terms separately. Using the uniform law of large numbers [e.g., Klyne and Shah, 2023, Lemma 9] on each component, it holds for all $\epsilon > 0$ that

$$\lim_{n_a\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\left\|\frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top - \mathbb{E}[Z_a Z_a^\top]\right\|_{\mathrm{op}} > \epsilon\right) = 0. \tag{4.D.5}$$

Moreover, we can expand the residual variance part as follows

$$\begin{aligned}\frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai}-\widehat{\pi}^\top Z_{ai})^2 = \frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai}-\pi^\top Z_{ai})^2 \\ + \frac{1}{\sqrt{n_a}}\left(\frac{2}{n_a}\sum_{i=1}^{n_a}(Y_{ai}-\pi^\top Z_{ai})\sqrt{n_a}(\widehat{\pi}-\pi)^\top Z_{ai}\right) \\ + \frac{1}{n_a}\left(\sqrt{n_a}(\widehat{\pi}-\pi)\left(\frac{1}{n_a}\sum_{i=1}^{n_a}Z_{ai}Z_{ai}^\top\right)\sqrt{n_a}(\widehat{\pi}-\pi)\right).\end{aligned}$$

Then, by the uniform asymptotic normality, the bounded moments of $Z$ and $Y$ and a further application of the law of large numbers [e.g., Klyne and Shah, 2023, Lemma 9] it follows for all $\epsilon > 0$ that

$$\lim_{n_a\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\left|\frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai}-\widehat{\pi}^\top Z_{ai})^2-\mathbb{E}[(u_a^Y)^2]\right|>\epsilon\right)=0. \tag{4.D.6}$$

Finally, denote $W_n \coloneqq \frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai}-\widehat{\pi}^\top Z_{ai})^2$, $W \coloneqq \mathbb{E}[(u_a^Y)^2]$, $V_n \coloneqq \frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top$ and $V \coloneqq \mathbb{E}[Z_a Z_a^\top]$. Then, by combining (4.D.5) and (4.D.6) it follows for all $\epsilon > 0$ that

$$\begin{aligned}&\sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_nV_n^{-1}-WV^{-1}\|_{\mathrm{op}}>\epsilon)\\ &\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_nV_n^{-1}-W_nV^{-1}\|_{\mathrm{op}}>\tfrac{\epsilon}{2})+\sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_nV^{-1}-WV^{-1}\|_{\mathrm{op}}>\tfrac{\epsilon}{2})\\ &\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_nV_n^{-1}-W_nV^{-1}\|_{\mathrm{op}}>\tfrac{\epsilon}{2})+\sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_nV^{-1}-WV^{-1}\|_{\mathrm{op}}>\tfrac{\epsilon}{2})\\ &\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P(\|V_n^{-1}-V^{-1}\|_{\mathrm{op}}\|W_n\|_{\mathrm{op}}>\tfrac{\epsilon}{2})+\sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_n-W\|_{\mathrm{op}}\|V^{-1}\|_{\mathrm{op}}>\tfrac{\epsilon}{2})\\ &\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P(\|V_n^{-1}-V^{-1}\|_{\mathrm{op}}\|W_n\|_{\mathrm{op}}>\tfrac{\epsilon}{2})+\sup_{P\in\mathcal{P}}\mathbb{P}_P(\|W_n-W\|_{\mathrm{op}}C>\tfrac{\epsilon}{2})\end{aligned}$$

By standard arguments and using the lower bound on the minimal eigenvalue of $V = \mathbb{E}[ZZ^\top]$ from Assumption 4.5.1, this proves (4.D.4) (left).

Combining the two convergence results in (4.D.4) and using that $n_a/n_b \to r$ shows that for all $\epsilon > 0$ it holds that

$$\lim_{\substack{n_a,n_b\to\infty\\ n_a/n_b\to r}}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi+\widehat{\Sigma}_\Pi(\beta))-(\tfrac{1}{r}\Sigma_\pi+\Sigma_\Pi(\beta))\|_{\mathrm{op}}>\epsilon\right)=0. \tag{4.D.7}$$

Furthermore, we can apply Johnson and Horn [1985, eq. (7.2.13)] to get that

$$\begin{aligned}&\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi+\widehat{\Sigma}_\Pi(\beta))^{1/2}-(\tfrac{1}{r}\Sigma_\pi+\Sigma_\Pi(\beta))^{1/2}\|_{\mathrm{op}}\\ &\le \|(\tfrac{1}{r}\Sigma_\pi+\Sigma_\Pi(\beta))^{-1/2}\|_{\mathrm{op}}\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi+\widehat{\Sigma}_\Pi(\beta))-(\tfrac{1}{r}\Sigma_\pi+\Sigma_\Pi(\beta))\|_{\mathrm{op}},\end{aligned}$$

which together with (4.D.7) and since Assumption 4.5.1 implies that $\inf_{P\in\mathcal{P}} \lambda_{\min}(\frac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta)) > 0$, implies for all $\epsilon > 0$ that

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{P\in\mathcal{P}} \mathbb{P}_P(\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} - (\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{1/2}\|_{\text{op}} > \epsilon) = 0.$$

Together with (4.D.3) this implies by Klyne and Shah [2023, Lemma 10 (b)] that

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^m} |\mathbb{P}_P\left(S_{n_a,n_b}(\beta) - \mu_{n_a,n_b}(\beta) \le t\right) - \Phi_m(t)| = 0,$$

where we in particular used that

$$\begin{aligned}
&S_{n_a,n_b}(\beta) - \mu_{n_a,n_b}(\beta) \\
&\quad = (\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta))^{-1/2}((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)) \\
&\quad = \Big((\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2}(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2}\Big)^{-1} \\
&\qquad\qquad \cdot \sqrt{n_b}(\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2}((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)) \\
&\quad = \Big(I + (\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2}\{(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} - (\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{1/2}\}\Big)^{-1} \\
&\qquad\qquad \cdot \sqrt{n_b}(\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2}((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)).
\end{aligned}$$

This completes the proof of Lemma 4.D.2. □

## 4.E. Experiment details and additional simulation results

### 4.E.1 Details of the simulated experiments in Section 4.4.1

**DGP1:** The individual-level data are generated from an SCM (4.2.1) with the following parameters

$$A := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad B := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0, \end{pmatrix}$$

$Z \overset{\text{iid}}{\sim} \mathcal{N}_m(0, I_m)$, $H \overset{\text{iid}}{\sim} \mathcal{N}_d(0, I_d)$ and $\nu^X, \nu^Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ with $g(H, \nu^X) := H + \nu^X$ and $h(H, \nu^Y) := H^\top \mathbf{1}_d + \nu^Y$. The true causal effect $\beta^* = (1, 2, 0, 0, 0)$.
**DGP2:** Let

$$A := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{pmatrix}, \text{Var}(Z) := \begin{pmatrix} 1 & 0.05 & -0.1 & 0.075 & 0.025 \\ 0.05 & 1 & 0 & 0 & 0 \\ -0.1 & 0 & 1 & 0 & 0 \\ 0.075 & 0 & 0 & 1 & 0 \\ 0.025 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$\mathrm{Var}(\nu^X) \coloneqq WW^\top + I_d$ where $W \in \mathbb{R}^{d\times d}$ with $W_i^j \overset{\text{iid}}{\sim} \mathrm{Unif}(-0.3, 0.5)$ for all $i, j \in [d]$, $\mathrm{Var}(\nu^Y) \coloneqq 1$, and $\mathrm{Cov}(\nu^X, \nu^Y) \in \mathbb{R}^{100}$ such that $\mathrm{Cov}(\nu^X, \nu^Y)^j$ is uniformly sampled from the set $\{0.2, 0.4, 0.6, 0.8\}$ for all $j \in \{1, \dots, 100\}$.

Then using $\beta^* \in \mathbb{R}^{100}$ with $(\beta^*)^1 \coloneqq 1$, $(\beta^*)^2 \coloneqq 2$, and $(\beta^*)^j \coloneqq 0$ for all $j \in \{3, \dots, 100\}$, we define $\Pi \coloneqq A^\top (I_d - B)^{-1}$ and $\pi \coloneqq \Pi\beta^*$. Moreover, based on the linear SCM and with $V \coloneqq (I_d - B)^{-1}\beta^*$ we have

$$\Sigma_\pi = \Big(V^\top \mathrm{Var}(\nu^X)V + \mathrm{Var}(\nu^Y) + 2V^\top \mathrm{Cov}(\nu^X, \nu^Y)\Big)\,\mathrm{Var}(Z)^{-1} \text{ and}$$
$$\Sigma_\Pi = \mathrm{Var}(\nu^X)^\top (I_d - B)^{-1} \otimes \mathrm{Var}(Z)^{-1}.$$

We then generated $\widehat{\pi}$, $\widehat{\Pi}$ from the following multivariate Gaussian distributions for a specific sample size $n$:

$$\widehat{\pi}_n \sim \mathcal{N}(\pi, \frac{1}{n}\Sigma_\pi) \qquad \text{and} \qquad \widehat{\Pi}_n \sim \mathcal{N}(\Pi, \frac{1}{n}\Sigma_\Pi).$$

### 4.E.2 Additional simulated experiments

We provide additional simulation results of a setting that the exclusion restriction criteria of IV is violated. The DGP is described below and the corresponding DAG is given in Figure 4.E.1.

**DGP3:** $m = 5$ and $d = 5$ and $\|\beta^*\|_0 = 2$. For increasing $n \coloneqq n_1 = n_2$, we generate iid $\{(Y_i, Z_i)\}_{i=1}^{n_1}$ and $\{(X_i, Z_i)\}_{i=1}^{n_2}$ according to the following SCM

$$\begin{aligned} X_i &\coloneqq AZ_i + BX_i + H_i + \nu_i^X \\ Y_i &\coloneqq X_i^\top \beta^* + Z_i^\top \gamma + H_i^\top \mathbf{1}_5 + \nu_i^Y, \end{aligned} \tag{4.E.1}$$

with the following parameters:

$$A = I_5, \quad B \coloneqq \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \quad \gamma = (0.1, 0.1), \quad \beta^* = (1, 2, 0, 0, 0),$$

$H_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_5)$, and $\nu_i^X, \nu_i^Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Then we compute the summary statistics using seemingly unrelated regression. The results are shown in Figure 4.E.2. The $\gamma$ parameter in (4.E.1) represents the violation of the exclusion restriction criteria. We see that as sample size goes larger, the bias and rmse continue to decrease. Although the Jaccard similarity and percentage of correct size start to decline, the average true positive rate (tpr) still stays around 100%. In this example, due to the invalid instruments, the estimated causal parents tend to be a superset of the true causal parent, but the estimated effects of the non-parent covaraites are relatively small.

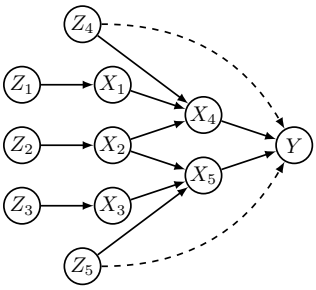



Figure 4.E.1: DAG for DGP3 which contains two invalid instruments violating the exclusion restriction criteria (dashed arrows).

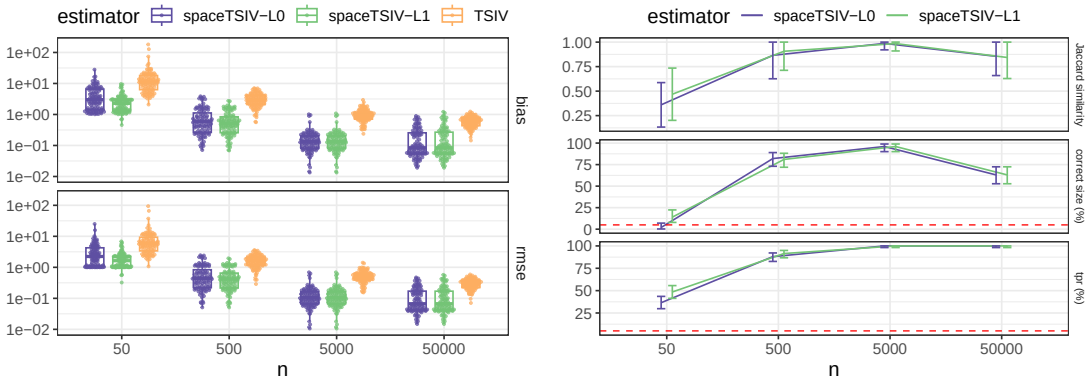



Figure 4.E.2: Results using data generated by DGP3 based on 100 repetitions. Left: Bias and rmse of the estimators. The y-axis is on log scale for clarity. Right: Average Jaccard similarity between the selected covariates and the true causal covariates (error bars indicate confidence intervals constructed by mean plus/minus one standard error), percentage of estimates that have the correct support size, and tpr (error bars indicate 95% binomial confidence intervals). This DGP contains 2 invalid instruments among the 5 instruments.

# Bibliography


E. Ailer, C. L. Müller, and N. Kilbertus. Instrumental Variable Estimation for Compositional Treatments. *arXiv Preprint arXiv:2106.11234*, 2024.

J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society B*, 44(2):139–160, 1982.

J. Aitchison. Principal Component Analysis of Compositional Data. *Biometrika*, 70(1): 57–65, 1983.

J. Aitchison. A General Class of Distributions on the Simplex. *Journal of the Royal Statistical Society B*, 47(1):136–146, 1985.

J. Aitchison and J. Bacon-Shone. Log Contrast Models for Experiments with Mixtures. *Biometrika*, 71(2):323–330, 1984.

J. Aitchison and M. Greenacre. Biplots of Compositional Data. *Journal of the Royal Statistical Society C*, 51(4):375–392, 2002.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

All of Us. URL https://allofus.nih.gov/, accessed: 2024-11-27, 2024.

T. W. Anderson and H. Rubin. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20 (1):46–63, 1949.

D. W. Andrews. Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica: Journal of the Econometric Society*, 61(4):821–856, 1993.

J. Angrist and G. Imbens. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

A. Aue and L. Horváth. Structural Breaks in Time Series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

J. Bai. Testing for Parameter Constancy in Linear Regressions: An Empirical Distribution Function Approach. *Econometrica: Journal of the Econometric Society*, 64(3): 597–622, 1996.

J. Bai. Estimating Multiple Breaks One at a Time. *Econometric Theory*, 13(3):315–352, 1997a.

J. Bai. Estimation of a Change Point in Multiple Regression Models. *Review of Economics and Statistics*, 79(4):551–563, 1997b.

J. Bai and P. Perron. Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica: Journal of the Econometric Society*, 66(1):47–78, 1998.

J. Bai and P. Perron. Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-Over-Threshold Detection of Multiple Change Points and Change-Point-Like Features. *Journal of the Royal Statistical Society B*, 1, 2019.

J.-M. Bardet, V. Brault, S. Dachian, F. Enikeeva, and B. Saussereau. Change-Point Detection, Segmentation, and Related Topics. *ESAIM: Proceedings and Surveys*, 68: 97–122, 2020.

F. Batool, A. Patel, D. Gill, and S. Burgess. Disentangling the Effects of Traits with Shared Clustered Genetic Predictors Using Multivariable Mendelian Randomization. *Genetic Epidemiology*, 46(7):415–429, 2022.

N. T. Baxter, M. T. Ruffin, M. A. Rogers, and P. D. Schloss. Microbiota-Based Model Improves the Sensitivity of Fecal Immunochemical Test for Detecting Colonic Lesions. *Genome Medicine*, 8(1):1–10, 2016.

Beijing Municipal Government. Beijing Heating Management Measures. https://www.beijing.gov.cn/gongkai/zfxxgk/zc/gz/202112/W020211216532156679225.pdf, 2009. Accessed: 2024-11-22.

A. S. Berry, K. Johnson, R. Martins, M. C. Sullivan, C. Farias Amorim, A. Putre, A. Scott, S. Wang, B. Lindsay, R. N. Baldassano, et al. Natural Infection with Giardia Is Associated with Altered Community Structure of the Human and Canine Gut Microbiome. *Msphere*, 5(4):e00670–20, 2020.

J. Bien, X. Yan, L. Simpson, and C. L. Müller. Tree-Aggregated Predictive Modeling of Microbiome Data. *Scientific Reports*, 11(1):1–13, 2021.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of Structural Causal Models with Cycles and Latent Variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

J. Bowden, F. Del Greco M, C. Minelli, Q. Zhao, D. A. Lawlor, N. A. Sheehan, J. Thompson, and G. Davey Smith. Improving the Accuracy of Two-Sample Summary-Data

Mendelian Randomization: Moving beyond the NOME Assumption. *International Journal of Epidemiology*, 48(3):728–742, 2019.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: Composable Transformations of Python+NumPy Programs, 2018. http://github.com/google/jax.

B. E. Brodsky and B. S. Darkhovsky. *Non-Parametric Methods in Change-Point Problems*. Kluwer Academic Publishers, 1993.

A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn. Compositional Data Analysis in the Geosciences: From Theory to Practice. In *GSL Special Publications*. Geological Society of London, 2006.

P. Bühlmann. Invariance, Causality and Robustness. *Statistical Science*, 35(3):404–426, 2020.

P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, 2011.

G. Cammarota, G. Ianiro, A. Ahern, C. Carbone, A. Temko, M. J. Claesson, A. Gasbarrini, and G. Tortora. Gut Microbiome, Big Data and Machine Learning to Promote Precision Medicine for Cancer. *Nature Reviews Gastroenterology & Hepatology*, 17 (10):635–648, 2020.

J. Chen and H. Li. Kernel Methods for Regression Analysis of Microbiome Compositional Data. In *Topics in Applied Statistics*, pages 191–201. Springer-Verlag, 2013.

S. Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5RK5G.

J. Cheng, J. Su, T. Cui, X. Li, X. Dong, F. Sun, Y. Yang, D. Tong, Y. Zheng, Y. Li, et al. Dominant Role of Emission Reduction in PM 2.5 Air Quality Improvement in Beijing during 2013–2017: a Model-Based Decomposition Analysis. *Atmospheric Chemistry and Physics*, 19(9):6125–6146, 2019.

G. C. Chow. Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrica: Journal of the Econometric Society*, 28(3):591–605, 1960.

R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. A Causal Framework for Distribution Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

P. L. Combettes and C. L. Müller. Regression Models for Compositional Data: General Log-Contrast Formulations, Proximal Optimization, and Microbiome Data Applications. *Statistics in Biosciences*, 13(2):217–242, 2021.

J. G. Cragg and S. G. Donald. Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, 9(2):222–240, 1993.

R. Davidson and J. G. MacKinnon. Confidence Sets Based on Inverting Anderson–Rubin Tests. *The Econometrics Journal*, 17(2):S39–S58, 2014.

A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002.

R. De La Cruz and J.-U. Kreft. Geometric Mean Extension for Data Sets with Zeros. *arXiv Preprint arXiv:1806.06403*, 2018.

P. J. Dhrymes. *Econometrics: Statistical Foundations and Applications*. Springer-Verlag, 2012.

M. Di Marzio, A. Panzera, and C. Venieri. Non-Parametric Regression for Compositional Data. *Statistical Modelling*, 15(2):113–133, 2015.

J.-M. Dufour. Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models. *Econometrica: Journal of the Econometric Society*, 65 (6):1365–1387, 1997.

C. Eckart and G. Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35 (3):279–300, 2003.

A. D. Fernandes, J. M. Macklaim, T. G. Linn, G. Reid, and G. B. Gloor. ANOVA-like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLOS One*, 8(7):e67019, 2013.

FinnGen. URL `https://www.finngen.fi/en`, accessed 2024-11-27, 2024.

J. Friedman and E. J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, 8(9):e1002687, 2012.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

P. Fryzlewicz. Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

P. Fryzlewicz. Narrowest Significance Pursuit: Inference for Multiple Change-Points in Linear Models. *Journal of the American Statistical Association*, 119(546):1633–1646, 2024.

V. Garfield, A. Salzmann, S. Burgess, and N. Chaturvedi. A Guide for Selection of Genetic Instruments in Mendelian Randomization Studies of Type 2 Diabetes and Hba1C: Toward an Integrated Approach. *Diabetes*, 72(2):175–183, 2023.

D. Gevers, S. Kugathasan, L. A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease.

G. Gibson. Rare and Common Variants: Twenty Arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8: 2224, 2017.

W. Gou, C.-W. Ling, Y. He, Z. Jiang, Y. Fu, F. Xu, Z. Miao, T.-Y. Sun, J.-S. Lin, H.-L. Zhu, L. Zhou, Y.-M. Chen, and J.-S. Zheng. Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated with Type 2 Diabetes. *Diabetes Care*, 44(2):358–366, 2021.

A. J. Grant and S. Burgess. Pleiotropy Robust Methods for Multivariable Mendelian Randomization. *Statistics in Medicine*, 40(26):5813–5830, 2021.

A. J. Grant and S. Burgess. An Efficient and Robust Approach to Mendelian Randomization with Measured Pleiotropic Effects in a High-Dimensional Setting. *Biostatistics*, 23(2):609–625, 2022.

GWAS Catalog. URL https://www.ebi.ac.uk/gwas/, accessed 2024-10-02, 2024.

M. Gönen and E. Alpaydın. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

T. Haavelmo. The Statistical Implications of a System of Simultaneous Equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12, 1943.

B. E. Hansen. Testing for Structural Change in Conditional Models. *Journal of Econometrics*, 97(1):93–115, 2000.

L. P. Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica: Journal of the Econometric Society*, 4(4):1029–1054, 1982.

F. P. Hartwig, N. M. Davies, G. Hemani, and G. Davey Smith. Two-Sample Mendelian Randomization: Avoiding the Downsides of a Powerful, Widely Applicable but Potentially Fallible Technique. *International Journal of Epidemiology*, 45(6):1717–1726, 2016.

D. M. Hawkins. Point Estimation of the Parameters of Piecewise Regression Models. *Journal of the Royal Statistical Society C*, 25(1):51–57, 1976.

M. Hein and O. Bousquet. Hilbertian Metrics and Positive Definite Kernels on Probability Measures. In *International Workshop on Artificial Intelligence and Statistics*, pages 136–143. PMLR, 2005.

G. Hemani, J. Bowden, and G. Davey Smith. Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Human Molecular Genetics*, 27(R2):R195–R208, 2018.

P. W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal Discovery from Heterogeneous/Nonstationary Data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

C. Huang, B. J. Callahan, M. C. Wu, S. T. Holloway, H. Brochu, W. Lu, X. Peng, and J.-Y. Tzeng. Phylogeny-Guided Microbiome OTU-Specific Association Test (POST). *Microbiome*, 10(1):1–15, 2022.

S. Huang, E. Ailer, N. Kilbertus, and N. Pfister. Supervised Learning and Model Analysis with Compositional Data. *PLOS Computational Biology*, 19(6):e1011240, 2023.

S. Huang, J. Peters, and N. Pfister. Causal Change Point Detection and Localization. *arXiv Preprint arXiv:2403.12677*, 2024a.

S. Huang, N. Pfister, and J. Bowden. Sparse Causal Effect Estimation Using Two-Sample Summary Statistics in the Presence of Unmeasured Confounding. *arXiv Preprint arXiv:2410.12300*, 2024b.

Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*, 486(7402):207–214, 2012.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

D. A. Jackson. Compositional Data in Community Ecology: The Paradigm or Peril of Proportions? *Ecology*, 78(3):929–940, 1997.

Japan Biobank. URL https://biobankjp.org/en/, accessed 2024-11-27, 2024.

C. R. Johnson and R. A. Horn. *Matrix Analysis*. Cambridge university press Cambridge, 1985.

F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed. Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control. *Nature*, 498(7452):99–103, 2013.

A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, 8:2114, 2017.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal Detection of Changepoints with a Linear Computational Cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

F. Kleibergen and R. Paap. Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics*, 133(1):97–126, 2006.

H. Klyne and R. D. Shah. Average Partial Effect Estimation Using Double Machine Learning. *arXiv Preprint arXiv:2308.09207*, 2023.

R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein. Best Practices for Analysing Microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

A. D. Kostic, D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, et al. Genomic Analysis Identifies Association of Fusobacterium with Colorectal Carcinoma. *Genome Research*, 22(2):292–298, 2012.

S. Kovács, H. Li, P. Bühlmann, and A. Munk. Seeded Binary Segmentation: a General Methodology for Fast and Optimal Changepoint Detection. *Biometrika*, 110(1):249–256, 2023.

J. Lafferty, G. Lebanon, and T. Jaakkola. Diffusion Kernels on Statistical Manifolds. *Journal of Machine Learning Research*, 6(1):129–163, 2005.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact Post-Selection Inference, with Application to the Lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

T. Leinster and C. Cobbold. Measuring Diversity: The Importance of Species Similarity. *Ecology*, 93:477–89, 2012.

F. Leonardi and P. Bühlmann. Computationally Efficient Change Point Detection for High-Dimensional Regression. *arXiv Preprint arXiv:1601.03704*, 2016.

B. Li, C. Yoon, and J. Ahn. Reproducing Kernels and New Approaches in Compositional Data Analysis. *Journal of Machine Learning Research*, 24(327):1–34, 2023.

H. Li. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

H. Lin and S. D. Peddada. Analysis of Microbial Compositions: a Review of Normalization and Differential Abundance Analysis. *NPJ Biofilms and Microbiomes*, 6(1):1–13, 2020.

W. Lin, P. Shi, R. Feng, and H. Li. Variable Selection in Regression with Compositional Covariates. *Biometrika*, 101(4):785–797, 2014.

M. Londschien and P. Bühlmann. Weak-Instrument-Robust Subvector Inference in Instrumental Variables Regression: A Subvector Lagrange Multiplier Test and Properties of Subvector Anderson-Rubin Confidence Sets. *arXiv Preprint arXiv:2407.15256*, 2024.

C. Lozupone and R. Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

A. R. Lundborg and N. Pfister. Perturbation-Based Analysis of Compositional Data. *arXiv Preprint arXiv:2311.18501*, 2023.

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The Projected Covariance Measure for Assumption-Lean Variable Significance Testing. *arXiv Preprint arXiv:2304.01098*, 2022.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, 2005.

J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, 35(3):253–278, 2003.

D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, et al. American Gut: An Open Platform for Citizen Science Microbiome Research. *Msystems*, 3(3):e00031–18, 2018.

C. Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

J. Morais and C. Thomas-Agnan. Impact of Covariates in Compositional Models and Simplicial Derivatives. *Austrian Journal of Statistics*, 50(2):1–15, 2021.

X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, et al. Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment. *Genome Biology*, 13(9): 1–18, 2012.

Y. S. Niu, N. Hao, and H. Zhang. Multiple Change-Point Detection: a Selective Overview. *Statistical Science*, 31(4):611–623, 2016.

L. Orváth and P. Kokoszka. Change-Point Detection with Non-Parametric Regression. *Statistics: A Journal of Theoretical and Applied Statistics*, 36(1):9–31, 2002.

E. Page. A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika*, 42(3/4):523–527, 1955.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100–115, 1954.

J. Park, C. Yoon, C. Park, and J. Ahn. Kernel Methods for Radial Transformed Compositional Data with Many Zeros. In *International Conference on Machine Learning*, pages 17458–17472. PMLR, 2022.

E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata. Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, 12(7):e1004977, 2016.

A. Patel, D. Gill, D. Shungin, C. S. Mantzoros, L. B. Knudsen, J. Bowden, and S. Burgess. Robust Use of Phenotypic Heterogeneity at Drug Target Genes for Mechanistic Insights: Application of Cis-Multivariable Mendelian Randomization to GLP1R Gene Region. *Genetic Epidemiology*, 48(4):151–163, 2024.

V. Paz, H. S. Dashti, S. Burgess, and V. Garfield. Selection of Genetic Instruments in Mendelian Randomisation Studies of Sleep Traits. *Sleep Medicine*, 2023.

J. Pearl. *Causality*. Cambridge University Press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. A Complete Generalized Adjustment Criterion. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 682–691, 2015.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

P. Perron, Y. Yamamoto, and J. Zhou. Testing Jointly for Structural Changes in the Error Variance and Coefficients of a Linear Regression Model. *Quantitative Economics*, 11(3):1019–1057, 2020.

M. Z. Pesenson, S. K. Suram, and J. M. Gregoire. Statistical Analysis and Interpolation of Compositional Data in Materials Science. *ACS Combinatorial Science*, 17(2):130–136, 2015.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society B*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

N. Pfister. Causality lecture notes. https://niklaspfister.github.io/download/lecture_notes_causality.pdf, 2024. Accessed on 2024-11-24.

N. Pfister and J. Peters. Identifiability of Sparse Causal Effects Using Instrumental Variables. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2022.

N. Pfister, P. Bühlmann, and J. Peters. Invariant Causal Prediction for Sequential Data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes. *Nature*, 490(7418):55–60, 2012.

N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J. M. Batto, S. P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S. D. Ehrlich, S. Zheng, and L. Li. Alterations of the Human Gut Microbiome in Liver Cirrhosis. *Nature*, 513(7516):59–64, 2014.

K. S. Ramirez, J. W. Leff, A. Barberán, S. T. Bates, J. Betley, T. W. Crowther, E. F. Kelly, E. E. Oldfield, E. A. Shaw, C. Steenbock, M. A. Bradford, D. H. Wall, and N. Fierer. Biogeographic Patterns in below-Ground Diversity in New York City's Central Park Are Similar to Those Observed Globally. *Proceedings of the Royal Society B: Biological Sciences*, 281(1795):20141988, 2014.

E. Ramon, L. Belanche-Muñoz, F. Molist, R. Quintanilla, M. Perez-Enciso, and Y. Ramayo-Caldas. kernInt: A Kernel Framework for Integrating Supervised and Unsupervised Analyses in Spatio-Temporal Metagenomic Datasets. *Frontiers in Microbiology*, 12:60, 2021.

T. W. Randolph, S. Zhao, W. Copeland, M. Hullar, and A. Shojaie. Kernel-Penalized Regression for Analysis of Microbiome Data. *The Annals of Applied Statistics*, 12(1): 540, 2018.

J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, et al. Vaginal Microbiome of Reproductive-Age Women. *Proceedings of the National Academy of Sciences of the United States of America*, 108(supplement_1):4680–4687, 2011.

J. M. Rees, A. M. Wood, F. Dudbridge, and S. Burgess. Robust Methods in Mendelian Randomization via Penalization of Heterogeneous Causal Estimates. *PLOS One*, 14 (9):e0222362, 2019.

J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. Balances: a New Perspective for Microbiome Analysis. *MSystems*, 3 (4):e00053–18, 2018.

D. Rothenhäusler, N. Meinshausen, P. B. hlmann, and J. Peters. Anchor Regression: Heterogeneous Data Meet Causality. *Journal of the Royal Statistical Society B*, 83(2): 215–246, 2021.

A. Ruaud, N. Pfister, R. E. Leya, and N. D. Youngbluta. Interpreting Tree Ensemble Machine Learning Models with endoR. *bioRxiv Preprint bioRxiv: 2022.01.03.474763v1*, 2022.

D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700. Springer-Verlag, 2019.

E. Sanderson and F. Windmeijer. A Weak Instrument F-Test in Linear IV Models with Multiple Endogenous Variables. *Journal of Econometrics*, 190(2):212–221, 2016.

E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, et al. Mendelian Randomization. *Nature Reviews Methods Primers*, 2(1):6, 2022.

J. D. Sargan. The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

I. J. Schoenberg. Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, 44, 1938.

B. Schölkopf, A. J. Smola, and F. Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT press, 2002.

P. Shi, A. Zhang, and H. Li. Regression Analysis for Microbiome Compositional Data. *The Annals of Applied Statistics*, 10(2):1019 – 1040, 2016.

I. Shpitser, T. VanderWeele, and J. M. Robins. On the Validity of Covariate Adjustment for Estimating Causal Effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, 2010.

L. Simpson, P. Combettes, and C. Müller. c-Lasso - a Python Package for Constrained Sparse and Robust Regression and Classification. *Journal of Open Source Software*, 6:2844, 2021.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

B. K. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7), 2011.

S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, et al. Tara Oceans: Towards Global Ocean Ecosystems Biology. *Nature Reviews Microbiology*, 18(8):428–445, 2020.

D. Tang, D. Kong, and L. Wang. The Synthetic Instrument: From Sparse Association to Sparse Causation. *arXiv Preprint arXiv:2304.01098*, 2023.

P. Thangavel, D. Park, and Y.-C. Lee. Recent Insights into Particulate Matter (PM2. 5)-Mediated Toxicity in Humans: An Overview. *International Journal of Environmental Research and Public Health*, 19(12):7511, 2022.

F. Topsøe. Jenson-Shannon Divergence and Norm-Based Measures of Discrimination and Variation. *Preprint*, 2003.

B. D. Topçuoğlu, N. A. Lesniak, M. Ruffin IV, J. Wiens, and P. D. Schloss. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *MBio*, 11(3):e00434–20, 2020.

C. Truong, L. Oudre, and N. Vayatis. Selective Review of Offline Change Point Detection Methods. *Signal Processing*, 167:107299, 2020.

M. Tsagris and G. Athineou. *Compositional: Compositional Data Analysis*, 2021. URL https://CRAN.R-project.org/package=Compositional. R package version 4.5.

M. T. Tsagris, S. Preston, and A. T. Wood. A data-based power transformation for compositional data. In *Proceedings of CoDaWork'11: 4th international workshop on Compositional Data Analysis, Egozcue, JJ, Tolosana-Delgado, R. and Ortego, MI (eds.) 2011*. CIMNE, 2011.

M. C. Tsilimigras and A. A. Fodor. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5):330–335, 2016.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, 2007.

UK Biobank. URL https://www.ukbiobank.ac.uk/, accessed 2024-10-02, 2024.

G. van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, 2009.

D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, et al. Quantitative Microbiome Profiling Links Gut Community Variation to Microbial Load. *Nature*, 551(7681):507–511, 2017.

P. Vangay, B. M. Hillmann, and D. Knights. Microbiome Learning Repo (ML Repo): A Public Repository of Microbiome Regression and Classification Tasks. *Gigascience*, 8 (5):giz042, 2019.

L. Y. Vostrikova. Detecting "disorder" in Multidimensional Random Processes. In *Doklady Akademii Nauk*, volume 259, pages 270–274. Russian Academy of Sciences, 1981.

D. Wang, Z. Zhao, K. Z. Lin, and R. Willett. Statistically and Computationally Efficient Change Point Localization in Regression Settings. *Journal of Machine Learning Research*, 22(1):11255–11300, 2021a.

J. Wang, Q. Zhao, J. Bowden, G. Hemani, G. Davey Smith, D. S. Small, and N. R. Zhang. Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments. *PLOS Genetics*, 17(6):1–24, 2021b.

S. Wang and H. Kang. Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization. *Biometrics*, 78(4):1699–1713, 2022.

J. H. Wells and L. R. Williams. *Embeddings and Extensions in Analysis*, volume 84. Springer-Verlag, 2012.

F. Wilcoxon. *Individual Comparisons by Ranking Methods*. Springer-Verlag, 1992.

N. Wilson, N. Zhao, X. Zhan, H. Koh, W. Fu, J. Chen, H. Li, M. C. Wu, and A. M. Plantinga. MiRKAT: Kernel Machine Regression-Based Global Association Tests for the Microbiome. *Bioinformatics*, 37(11):1595–1597, 2021.

J. Witte, L. Henckel, M. H. Maathuis, and V. Didelez. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020. URL http://jmlr.org/papers/v21/20-175.html.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2010.

T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, et al. Human Gut Microbiome Viewed across Age and Geography. *Nature*, 486(7402):222–227, 2012.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. Strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7: 1–38, 2002.

S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen. Cautionary Tales on Air-Quality Improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

N. Zhao, J. Chen, I. M. Carroll, T. Ringel-Kulka, M. P. Epstein, H. Zhou, J. J. Zhou, Y. Ringel, H. Li, and M. C. Wu. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*, 96(5):797–807, 2015a.

N. Zhao, J. Chen, I. M. Carroll, T. Ringel-Kulka, M. P. Epstein, H. Zhou, J. J. Zhou, Y. Ringel, H. Li, and M. C. Wu. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*, 96(5):797–807, 2015b.

Q. Zhao, J. Wang, W. Spiller, J. Bowden, and D. S. Small. Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples. *Statistical Science*, 34(2):317–333, 2019.

Q. Zhao, J. Wang, G. Hemani, J. Bowden, and D. S. Small. Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score. *The Annals of Statistics*, 48(3):1742–1769, 2020.

Y.-H. Zhou and P. Gallins. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in Genetics*, page 579, 2019.

O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for Missing Heritability: Designing Rare Variant Association Studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):E455–E464, 2014.