Jeffrey Adams

Causal Inference and Causal Discovery with Latent Variables

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

February 2024

Jeffrey Adams
ja@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Causal Inference and Causal Discovery with Latent Variables
Supervisor:	Professor Niels Richard Hansen University of Copenhagen
Assessment Committee:	Associate Professor Niklas Pfister (chair) University of Copenhagen
	Professor Mathias Drton Technical University of Munich
	Emilie Devijver, CNRS Researcher Laboratoire d'Informatique de Grenoble
Date of Submission:	February 29, 2024
Date of Defense:	May 14, 2024
ISBN:	978-87-7125-226-2

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen. It received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 801199, and from a research grant (NNF20OC0062897) from Novo Nordisk Fonden.

Preface

The results in this PhD thesis were produced under the supervision of Professor Niels Richard Hansen at the Department of Mathematical Sciences, University of Copenhagen.

I am especially grateful to Niels for his encouragement during the easiest parts of the PhD, for his levelheadedness during the hardest parts of the PhD, for his precision and creativity during the most mathematical parts of the PhD, and for all the time he invests in his students and our projects.

I am also grateful to Kun Zhang for his continued mentorship since supervising my MS thesis, for being such a welcoming host in Abu Dhabi, and for his never-ending enthusiasm for causality research.

I would like to thank the rest of the Copenhagen Causality Lab for the many helpful and interesting conversations, and other colleagues and friends at Math for sharing in both the more fun and the less fun parts of doing a PhD.

Finally, I am forever grateful to my parents and grandparents, who worked long hours at hard jobs so that their children and grandchildren could do fun things like PhDs in Statistics.

Jeffrey Adams
February, 2024

Abstract

This thesis contains various results regarding the identifiability of causal models and estimation of causal effects in the presence of latent variables.

First, we study the identifiability of the partially observed linear causal model. In many applications, statistical dependencies between measured variables are partially due to unmeasured confounders or mediators. In order to fully understand the data generating process, it is desirable to learn not only the causal relations between the observed variables, but also the causal structure of the latent variables. To this end, we present two local graphical conditions that are necessary and sufficient to ensure identifiability of the full graph.

Second, we study the deconfounder algorithm, which was proposed for multiple causal inference in the presence of unmeasured confounding. The deconfounder can be seen as outcome regression adjusted for a substitute confounder which is recovered from the observed treatments. We give theoretical results justifying the use of this method when the treatments are independent when conditioning on the confounter. We also analyze the finite sample error of this estimator in terms of the recovery error of the confounder. The deconfounder is analyzed both from a causal and a causally agnostic perspective.

Third, we study the steady-state distributions of Lévy-driven Ornstein-Uhlenbeck process. We argue that the steady-state interventional distributions of these processes can be expressed in terms of the first two moments of the observational steady-state distribution under a condition we refer to as drift-volatility balance. We derive equations relating higher-order cumulants of the steady-state distributions to the parameters of the stochastic process. From the second- and third-order equations, we derive a rank constraint which holds when drift-volatility balance is satisfied.

Sammenfatning

Denne afhandling indeholder forskellige resultater vedrørende identificerbarheden af kausale modeller og estimering af kausale effekter i tilstedeværelsen af latente variabler.

Først studerer vi identificerbarheden af den delvist observerede lineære kausale model. I mange applikationer skyldes statistiske afhængigheder mellem målte variable delvist umålte confoundere eller mediatorer. For fuldt ud at forstå datagenereringsprocessen er det ønskeligt at lære ikke kun årsagssammenhængene mellem de observerede variabler, men også årsagsstrukturen af de latente variable. Til dette formål præsenterer vi to lokale grafiske forhold, der er nødvendige og tilstrækkelige til at sikre identificerbarheden af den fulde graf.

For det andet studerer vi deconfounder-algoritmen, som blev foreslået til multipel kausal inferens i nærvær af umålt confounding. Deconfounderen kan ses som udfaldsregression justeret for en substitutconfounder, som rekonstrueres fra de observerede folklarende variable. Vi giver teoretiske resultater, der retfærdiggør brugen af denne metode, når de forklarende variable er uafhængige, når de betinges på confounteren. Vi analyserer også den endelige prøvefejl i denne estimator i form af rekonstrueringsfejlen for confounderen. Deconfounderen analyseres både ud fra et kausalt og et kausalt agnostisk perspektiv.

For det tredje studerer vi steady-state-fordelingerne af den Lévy-drevne Ornstein-Uhlenbeck-proces. Vi viser, at de steady-state interventionelle fordelinger af disse processer kan udtrykkes i form af de første to momenter af den observationelle steady-state fordeling under en tilstand, vi omtaler som drift-volatilitetsbalance. Vi udleder ligninger, der relaterer højere-ordens kumulanter af steady-state distributioner til parametrene for den stokastiske proces. Fra anden- og tredjeordens ligninger udleder vi en rangbetingelser, som gælder, når drift-volatilitetsbalancen er opfyldt.

Contributions and Structure

Chapter 1 is an introduction to the thesis. It is not a literature review, but rather attempts to identify common threads between chapters and situate them in a common framework. More thorough reviews relating each of the thesis's contributions to existing work appear in the corresponding chapters.

The introduction is followed by 3 chapters, each containing a paper as well as additional results or discussion. For reference within this thesis, we give each paper an acronym, for example [Identification]. All theorems, etc., are numbered relative to the paper they appear in.

Chapter 2 (Identifiability in Partially Observed Linear Models) discusses the identifiability of a causal adjacency matrix when only a subset of causal variables are observed. The chapter contains the following paper:

[Identification] [Adams et al., 2021]. J. Adams, N. Hansen, and K. Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34: 22822–22833, 2021.

Chapter 3 (Multiple Causal Inference and Substitute Adjustment) discusses the deconfounder as a method for multiple causal inference. We develop theory for the method as a specific instance of a causally agnostic adjustment problem. Further, we relate the causally agnostic analysis of the method to a causal interpretation. The chapter contains the following paper:

[Adjustment] [Adams and Hansen, 2024]. J. Adams and N. R. Hansen. Substitute adjustment via recovery of latent variables. *arXiv preprint arXiv:2403.00202*, 2024.

Paper status: Submitted at JMLR.

Chapter 4 (Causal Interpretations of Lévy-Driven Ornstein Uhlenbeck Processes) studies steady-state distributions of Lévy driven Ornstein Uhlenbeck processes and a condition under which causal conclusions can be inferred from the observational distributions. The chapter consists of the following paper:

[Precision] [Recke et al., 2024]. C. O. Recke, J. Adams, and N. R. Hansen. Non-Gaussian graphical precision models. 2024.

Paper status: Work in progress.

Contents

1 Introduction

In this PhD thesis, we present theory for causal inference and causal discovery in the presence of latent variables. Roughly speaking, causal inference is concerned with the estimation of causal effects when the causal graph (i.e. the causal relations between variables) are known, whereas causal discovery (also known as causal search or structure learning) is concerned with the process of learning that structure itself [Spirtes et al., 2000, Pearl, 2009, Peters et al., 2017].

One famous problem in causal inference is the attempt to estimate the effect of a treatment (say, a drug) on a response (say, a disease state some time after treatment) in the presence of confounding factors (say, the severity of the disease state before treatment). If the treatment is systematically administered to sicker patients, it may seem to be less effective than the control regardless of its actual effectiveness because sicker patients are just more likely to stay sick. One goal of causal inference is to identify the treatment effect by removing the confounding bias.

Notice that in the above example, the causal order is known and corresponds to the temporal order. Often this is not the case. For example, in cellular biology, there are gene-regulatory networks which constitute causal systems; one gene up- or down-regulates another's expression, but it is not know a priori which direction this regulation takes place. The goal of causal discovery can be to learn the causal order between two variables, or to describe the causal structure of a large causal network.

We can go further. In both examples, we were primarily concerned (at least implicitly) with the causal effects between measured variables; any unobserved confounding was a nuisance to be removed. But we may also be interested in the unobserved part of the causal graph in its own right. Not only can latent structure help us explain the distribution of the observed variables, but it can also point to real causal variables whose existence can be hypothesized and experimentally confirmed.

The remainder of this introduction identifies two unifying themes throughout the thesis. In Section 1.1, we formalize the word "causal" in terms of structural equation models, with attention to how each of the papers included in this thesis fit into the structural equation framework. In Section 1.2, we discuss the relevance of non-Gaussianity and tensor decomposition algorithms to identification and estimation problems in causal discovery and causal inference.

1.1 Structural Equation Models

While "causation" may be a philosophically loaded term, it is supposed to be distinguished from mere conditioning by describing something "fundamental" about the data

generating process. While this can be given a mathematically precise meaning using the languages of interventions, causal graphs, or potential outcomes, here we present structural equation models as a formalization of causal relations. When interpreted causally, structural equation models describe the mechanisms by which some variables (causally) influence other variables, thereby inducing causal graphs, interventional distributions, and distributions over potential outcomes [Peters et al., 2017, for example].

A structural equation model over variables $X_1, ..., X_p$ is a model of the form

$$X_i = f_i(\text{Pa}(X_i), \varepsilon_i) \tag{1}$$

where f_i is an arbitrary function in a function class $\mathcal{F}$, where $\text{Pa}(X_i) \subseteq \{X_1, ..., X_p\}$, and where $\{\varepsilon_1, ..., \varepsilon_p\}$ are unmeasured jointly independent random variables. We usually refer to $\text{Pa}(i)$ as the parents of X_i, and ε_i as the independent noise.

Together, the equations of (1) along with the distributions of ε induce a distribution over X—we call this the **observational distribution**. If the "=" sign is interpreted as mere equality in distribution, the (1) makes no causal claim. However, the "=" sign can also be interpreted as value assignment, so that in addition to describing the joint distribution of $X_1, ..., X_p$, it also induces interventional distributions [Pearl, 2009]. In this case, it is common to refer to the structural equation model as a structural *causal* model, and to refer to f_i as a *causal* mechanism.

One line of research in causal inference (when $\text{Pa}(X_i)$ are known) and causal discovery (when they are not) is concerned with the identification and estimation of structural equation models from the observational distribution, which may be possible under appropriate restrictions on f_i and the distributions of ε. For example, identification of the mechanism and noise distributions is possible in additive [Shimizu et al., 2006, Peters et al., 2013] and post non-linear additive [Zhang and Hyvärinen, 2009, Qiao et al., 2021] noise models. This is not an inherently causal endeavor; f_i are just estimands, and estimands don't have to be causal to be interesting. However, if the modeling assumptions (here, independent noise, the function class $\mathcal{F}$, and in causal inference the order of the variables) are plausible approximations of the true data generating process, we may regard the structural equation model as a structural causal model.

Let's analyze one particular structural causal model for the confounding example from the beginning of this chapter. Writing X for the treatment, Y for the response, and Z for the confounder, the example was implicitly written with $X \in \mathbb{R}$. However, in Chapter 3 we consider the case where $X \in \mathbb{R}^p$ (so that there are p treatments), and where the treatments $X_1, ..., X_p$ are mutually independent conditional on the confounder Z. Expressed as a partially linear structural causal model,

$$\begin{aligned} Z &= \varepsilon_z \\ X_i &= f_i(Z, \varepsilon_i), \quad i \in \{1, ..., p\} \\ Y &= \sum_{i=1}^{p} \beta_i X_i + f_z(Z) + \varepsilon_y. \end{aligned} \tag{2}$$

Here β_i are the parameters of interest, representing the linear effect of X on Y. It is well known that if Z is unmeasured, then β is not identifiable from the joint distribution of (X, Y) without additional assumptions. Nevertheless, under appropriate assumptions on the conditional distributions $X|Z = z$, the σ-algebra generated by Z is identifiable from the joint distribution of X. In such case, β is also identified in (2). Chapter 3 discusses the recovery of Z and the estimation of β when such assumptions are approximately satisfied.

Another famous structural equation model is the linear non-Gaussian acyclic model (LiNGAM) Shimizu et al. [2006]. Writing $X = (X_1, ..., X_p)^T$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_p)^T$,

$$X = \mathbf{F}X + \varepsilon, \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{p \times p}$ is an **adjacency matrix**, where at most one ε_i is Gaussian and none are degenerate, and where there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{PFP}^T$ is lower triangular. The appropriateness of "linear" and "non-Gaussian" are self-evident. To see why they are also called acyclic, notice that X_i depends only on the support of the i-th row of $\mathbf{F}$, so that this support is equal to $\mathrm{Pa}(X_i)$. The requirement that $\mathbf{F}$ be lower triangular (modulo permutation) indicates that Pa is a partial ordering (i.e. an acyclic relation) on $\{X_1, ..., X_p\}$. Causally speaking, the interventional distribution fixing $X_i = x$ is expressed by setting the i-th row of $\mathbf{F}$ to zero and $\varepsilon_i = x$ in (3).

Solving in terms of the observed variables X, we have

$$X = \mathbf{M}\varepsilon \tag{4}$$

where $\mathbf{M} := (\mathbf{I} - \mathbf{F})^{-1}$ is called the **mixing matrix**. Notice that (3) and (4) entail the same observational distribution, but represent different structural causal models; in the latter, no X_i causes any other X_j. However, under either causal model, the (i, j)-th slot of $\mathbf{M}$ represents the net effect of ε_j on X_i. We will never regard (4) as causal in this thesis.

In general, the problem of identifying $\mathbf{M}$ and the distributions of ε is known as independent component analysis (ICA) Comon [1994], Hyvärinen et al. [2001]. Clearly, $\mathbf{M}$ is at best identifiable up to scaling and permutation of columns; this corresponds to reindexing and rescaling the unobserved signals ε. As we will see in Section 1.2, it turns out that under the conditions of LiNGAM, $\mathbf{M}$ is always identifiable up to permutation and scaling of columns from the joint distribution of X in the case where all of $X_1, ..., X_p$ from (3) are observed. From this fact, Shimizu et al. [2006] constructively prove the identifiability of (3) in the case where all variables X are observed and their order is unknown. In Chapter 2, we study the same problem in the partially observed case, so that we observe only only a subset of $X_1, ..., X_p$.

When interpreted causally, structural equation models of the form (1) often indicate that causes do not change their values after they have brought about their effects. Alternatively, in the case of (3), one observation of X can be interpreted as a single draw of ε propagated through the causal network in continuous time, and then measured after the system has reached equilibrium. (An analogous result was originally shown in discrete

time by Fisher [1970].) Formally, this entails a differential equation of the form

$$dX(t) = (\mathbf{F} - \mathbf{I})X(t)dt + \varepsilon dt \tag{5}$$

where $\varepsilon \in \mathbb{R}^p$ is constant across time (but random across observations). The solution to (5) is given by

$$X(t) = e^{t(\mathbf{F}-\mathbf{I})}X(0) - \int_0^t e^{t(\mathbf{F}-\mathbf{I})}\varepsilon dt. \tag{6}$$

When $\mathbf{F}$ is lower triangular as in LiNGAM, or more generally when all its eigenvalues have real part strictly less than one, it is the case that

$$\lim_{t\to\infty} X(t) = (\mathbf{I} - \mathbf{F})^{-1}\varepsilon = \mathbf{M}\varepsilon \tag{7}$$

regardless of the initial value of $X(0)$. Hence the equilibrium distribution of (5) coincides with the observational distribution of (3).

Furthermore, if a causal interpretation of (5) is desired, the intervention fixing $X_i(t) = x$ across all times t corresponds to setting the i-th row of $\mathbf{F}$ to zero and $\varepsilon_i = x$; hence the equilibrium interventional distribution of (5) and the interventional distrubition of (3) coincide. Thus while (5) is not formally a structural equation model of the form (1) (since it references infinitely many intermediate values of $X(t)$) it nevertheless has a clear causal interpretation.

Interestingly, (6) is an Ornstein-Uhlenbeck process with the Brownian motion $dZ(t)$ replaced by a constant drift process εdt. In Chapter 4, we study a generalization of (5). Rather than time-constant ε, we allow a time-varying semi-martingale $Z(t)$:

$$dX(t) = (\mathbf{F} - \mathbf{I})X(t)dt + dZ(t). \tag{8}$$

In this case, unlike (5), the system will never reach equilibrium in the strict sense that $X(t) \to X$. However, so long as $Z(t)$ is a Lévy process and all eigenvalues of $\mathbf{F}$ have real part strictly less than one, a steady-state distribution exists and is given by

$$X = \lim_{t\to\infty} \int_0^t e^{s(\mathbf{F}-\mathbf{I})}dZ(s). \tag{9}$$

Notice that when $Z(t)$ is the pure drift Lévy process εt, this is precisely the limit of (6). Chapter 4 provides a causal interpretation of Lévy-driven Ornstein-Uhlenbeck processes by describing their steady-state distributions under time-constant interventions.

1.2 Non-Gaussianity and Tensor Methods

A recurring theme throughout this thesis is the blessing of non-Gaussianity due to the generic uniqueness of tensor decompositions.

Consider the representation of the LiNGAM model given by (4). We are interested in identifying $\mathbf{M}$; from this, the adjacency matrix $\mathbf{F} = \mathbf{M}^{-1} + \mathbf{I}$ is also identifiable. Due to

mutual independence of ε, the covariance of X is given by

$$\text{Cov}(X) = \mathbf{M}\,\text{Cov}(\varepsilon)\mathbf{M}^T = \sum_{i=1}^{p} \text{Var}(\varepsilon_i)(\mathbf{M}e_i) \otimes (\mathbf{M}e_i)$$

where e_i denotes the i-th standard basis vector. It is well known that a symmetric matrix never has a unique decomposition into a sum of symmetric rank-one matrices unless it is itself rank one. However, consider instead the third cumulant of X [**?**]:

$$\text{cum}_3(X) = \sum_{i=1}^{p} \text{cum}_3(\varepsilon_i)(\mathbf{M}e_i) \otimes (\mathbf{M}e_i) \otimes (\mathbf{M}e_i). \tag{10}$$

It turns out that this and many similar three-way tensors often do have a unique decomposition into a sum of p rank-one tensors due to Kruskal's theorem.

Define the **Kruskal rank** of a matrix as the largest number r such that every set of r columns is linearly independent.

Theorem 1 (Kruskal)**.** *Let $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ each have q columns. Suppose that the sum of their Kruskal ranks is greater than $2q + 1$. Suppose $\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2, \tilde{\mathbf{M}}_3$ also have q columns, and that*

$$\sum_{i=1}^{q} (\mathbf{M}_1 e_i) \otimes (\mathbf{M}_2 e_i) \otimes (\mathbf{M}_3 e_i) = \sum_{i=1}^{q} (\tilde{\mathbf{M}}_1 e_i) \otimes (\tilde{\mathbf{M}}_2 e_i) \otimes (\tilde{\mathbf{M}}_3 e_i). \tag{11}$$

Then there exist a permutation matrix $\mathbf{P}$ and invertible diagonal matrices $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ with $\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$ such that $\tilde{\mathbf{M}}_i = \mathbf{M}_i\mathbf{D}_i\mathbf{P}$ for each $i \in \{1, 2, 3\}$.

Theorem 1 was originally proven in Kruskal [1977]; for a more concise proof, see Rhodes [2010]. In words, (1) provides a condition under which a tensor T is uniquely decomposable into a sum of rank one tensors. Applied to (10), if $\text{cum}_3(\varepsilon_i) \neq 0$, then these conditions are always sufficient for identification of $\mathbf{M}$ up to permutation and scaling of columns; the mixing matrix $\mathbf{M}$ always has full Kruskal rank when $\mathbf{F}$ is strictly lower triangular. Given the columns of $\mathbf{M}$ (or estimates thereof), the main contribution of Shimizu et al. [2006] is an algorithm to identify (or to estimate) $\mathbf{F}$.

Notice that for i with $\text{cum}_3(\varepsilon_i) = 0$, the triple product of $\mathbf{M}e_i$ vanishes from (10), and so is not identifiable by Kruskal's theorem. However, if $\text{cum}_k(\varepsilon_i) \neq 0$ for some $k > 2$, then for this k, Kruskal's theorem identifies $\mathbf{M}e_i$ from some three-way subtensor of $\text{cum}_k(X)$. We now see the importance of non-Gaussianity in linear acyclic models.

Proposition 1. *Let X be a random variable. Then $\text{cum}_k(X) = 0$ for all $k > 2$ if and only if X is Gaussian.*

Theorem 1 also reveals the mathematical advantages of using of finite mixture models in (2). If X is a finite mixture of product measures, then for all $i \neq j \neq k \neq i$ it is the

case that

$$\mathbb{E}[X_i X_j X_k] = \sum_{z=1}^{m} \mathbb{P}[Z = z]\mathbb{E}[X_i|Z = z]\mathbb{E}[X_j|Z = z]\mathbb{E}[X_k|Z = z] \tag{12}$$

by mutual independence of X given Z. It follows that for every non-empty $I, J, K \subset \{1, ..., p\}$, if I, J, K are disjoint, then

$$\mathbb{E}(X_I \otimes X_J \otimes X_K) = \sum_{z=1}^{m} \mathbb{P}[Z = z]\mathbb{E}[X_I|Z = z] \otimes \mathbb{E}[X_J|Z = z] \otimes \mathbb{E}[X_K|Z = z] \tag{13}$$

which has the form of the sum in (11). Under appropriate conditions on the conditional means, the latter are identifiable from the raw third moment $\mathbb{E}[X \otimes X \otimes X]$. Indeed **?** provide conditions under which each conditional probability $X_i|Z = z$ is identifiable from the joint distribution of X using a similar observation.

Unfortunately, neither Kruskal's theorem nor the identifiability results of **?** are constructive, and various heuristics must be used. Moreover, when the relevant cumulant tensors are estimated rather than known, any tensor decomposition algorithm must be able to handle small errors in the tensor's estimation. For example, Anandkumar et al. [2014] and Guo et al. [2022] provide probably approximately correct decomposition algorithms when (13) holds for some I, J, K; we discuss these algorithms further in Chapter 3. While these algorithms are also applicable to tensors of the form in (10), it is common to use computationally faster algorithms such as FastICA **?** when (10) denotes a $p \times p \times p$ tensor with p rank-one addends. However, in the case where there are more than p rank-one components in the tensor sum—as in the partially observed case—none of the aforementioned algorithms apply. We review alternative methods in Chapter 2.

Proposition 1 also motivates the study of Ornstein-Uhlenbeck processes (8) driven by non-Brownian Lévy processes. In Chapter 4, we show that

$$\operatorname{cum}_k(Z(1)) + \sum_{i=1}^{k} \mathbf{M} \times_i \operatorname{cum}_k(X) = 0 \tag{14}$$

for all k; the case where $k = 2$ is the continuous-time Lyapunov equation. If $Z(t)$ is a Brownian motion, then both Z and X are Gaussian, so that every term of (14) is trivially zero for $k > 2$ due to Proposition 1. However, if $Z(t)$ is any other (non-deterministic) Lévy process, then (14) is non-trivial for some k.

Under a condition we call rank-volatility balance, Chapter 4 shows that interventional steady-state distributions of X can be expressed in terms of the observational distribution of X. Furthermore, we use the second- and third-order cumulant equations to construct a matrix $\mathbf{V}$ which is rank deficient under that same condition. Therefore, Chapter 4 proposes to turn the singular values of $\mathbf{V}$ into a test for rank volatility balance. However, if X is Gaussian, then $\mathbf{V}$ is identically 0 whether drift-volatility balance is satisfied or not, and so the test is useless in the Gaussian case.

2 Identifiability in Partially Observed Linear Models

This chapter contains the following paper:

[Identification] [Adams et al., 2021]. J. Adams, N. Hansen, and K. Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34: 22822–22833, 2021.

Here we consider the identifiability of an adjacency matrix from the (scaled and permuted) columns of the mixing matrix in partially observed linear causal models. In Section 6 of the paper, we argue that our model assumptions unify a wide variety of related partially observed causal models, and that the assumptions of many of those models are special cases of our conditions.

In order to discuss identifiability in this model class, Section 2 of the paper motivates and leverages a condition that we refer to as minimality (of the edges in the adjacency matrix). However, in the original paper, the two notions of minimality and identifiability are sometimes conflated. Stated clearly, Theorem 2 proves that every minimal admissible adjacency matrix satisfies two structural assumptions. Theorem 3 provides a constructive proof that the true adjacency matrix can be recovered from the columns of the true mixing matrix whenever the structural conditions are satisfied.

After the paper, in Section 2.1, we will discuss various generalizations of the notion of identifiability with reference to the theoretical results of the main paper. In light of these, we present an additional theoretical result regarding the identifiability of partially observed linear models.

Identification of Partially Observed Linear Causal Models: Graphical Conditions for the Non-Gaussian and Heterogeneous Cases

Jeffrey Adams[1]*, **Niels Richard Hansen**[1], **Kun Zhang**[2]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark
[2]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
ja@math.ku.dk, niels.r.hansen@math.ku.dk, kunz1@cmu.edu

Abstract

In causal discovery, linear non-Gaussian acyclic models (LiNGAMs) have been studied extensively. While the causally sufficient case is well understood, in many real applications the observed variables are not causally related. Rather, they are generated by latent variables, such as confounders and mediators, which may themselves be causally related. Existing results on the identification of the causal structure among the latent variables often require very strong graphical assumptions. In this paper, we consider partially observed linear models with either non-Gaussian or heterogeneous errors. In that case we give two graphical conditions which are necessary for identification of the causal structure. These conditions are closely related to sparsity of the causal edges. Together with one additional condition on the coefficients, which holds generically for any graph, the two graphical conditions are also sufficient for identifiability. These new conditions can be satisfied even when the number of latent variables is very large. We demonstrate the validity of our results on synthetic data.

1 Introduction

In the standard causal discovery problem, we are given non-experimental data and aim to learn the direct causal relations between the observed variables [1, 2]. But in many applications, we do not believe that all causal variables relevant to the observed system have been measured. While some of the observed variables may interact directly, others might interact indirectly via latent mediators, and still others could be generated by latent common causes; indeed, any pair of observed variables may stand in all three relations at once. Further, the relevant latent variables may be causally related themselves. For example, responses to psychometric questionnaires are usually thought of as noisy views of various traits, and the researcher is predominately interested in the causal relations between these hidden traits and their hierarchical structure. Similarly, in financial markets, stock returns may be causally related, but may also be confounded or mediated by a complicated network of unmeasured economic and political factors.

It is natural to ask what conditions are both necessary and sufficient for the identification of such partially observed causal structures from observational data. Various sufficient conditions have been proposed; however, these conditions are rather restrictive, and are not in general necessary for identification of the full causal structure.

*The work presented in this article was started while JA was at CMU.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In this work, we consider the case of linear causal models in which the overcomplete mixing matrix from the noise terms to the measured variables is identifiable up to permutation and scaling of columns. This is possible, for example, in the case of independent non-Gaussian noise [3], or when given access to heterogeneous domains in which the variances of the noise terms change independently across domains but the causal graph and weights remain constant (see Theorem 1 of our paper). We provide necessary and sufficient conditions under which the latent causal structure can be uniquely identified up to trivial indeterminicies.

2 Problem setup

Suppose some causal variables $\mathcal{V} = \{V_1, \ldots, V_p\}$ follow a linear structural equation model (SEM)

$$\mathbf{V} = \mathbf{F}\mathbf{V} + \varepsilon, \tag{1}$$

where $\mathbf{V} := (V_1, \ldots, V_p)^T$ is a vector of causal variables, $\mathbf{F}$ is a causal adjacency matrix that can be permuted (by simultaneous row and column permutations) to strictly lower-triangular form, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^T$ is a vector of independent noise variables. In this paper we consider two settings for ε_i: 1) all ε_i are mutually independent and non-Gaussian; or 2) there are multiple domains, ε_i are uncorrelated within each domain, and their variances change independently across domains. We will make the second assumption technically precise Section 3.

We seek necessary and jointly sufficient conditions for identifiability of $\mathbf{F}$ (up to trivial indeterminacies) in the case where only some subset of $\mathcal{V}$ (which we call $\mathcal{X}$) is measured. Thus $\mathbf{F}$ may encode observed-observed interactions, latent confounding, latent-latent interactions, and latent mediation or intermediate confounding. Our results identify $\mathbf{F}$ from the equivalence class $\mathcal{M}$, as defined in Section 2.2, of mixing matrices induced by (1). This equivalence class $\mathcal{M}$ is identifiable if, for example, the errors are non-Gaussian or if their distribution changes over time or between domains.

2.1 Notation

For any matrix $\mathbf{A}$ and index sets J and K, we write $\mathbf{A}_K^J$ to denote the submatrix of $\mathbf{A}$ with columns indexed by J and rows indexed by K. Observe that $\mathbf{A}_K^J = \mathbf{I}_K \mathbf{A} \mathbf{I}^J$. Thus, $\mathbf{F}_K^J$ describes the direct effect of $\{V_j : j \in J\}$ on $\{V_k : k \in K\}$. (Remember: causes are *up*-stream of their effects.)

The graph induced by (1) has edges $V_i \to V_j$ whenever $\mathbf{F}_j^i \neq 0$. We write $\mathrm{Pa}(V_i) := \{V_j : V_i \leftarrow V_j\}$ and $\mathrm{Ch}(V_i) := \{V_j : V_i \to V_j\}$, respectively, to denote the parents and children of V_i. We say that $(V_1, ..., V_k)$ constitutes a **directed path** from V_1 to V_k if $V_i \to V_{i+1}$ for every $i \in \{1, ..., k-1\}$. Trivially, for every V_i, (V_i) is a directed path from V_i to itself; we say that a directed graph is **acyclic** (a DAG) if (V_i) is the only such path. We write $\mathrm{Anc}(V_i) := \{V_j \in \mathcal{V} : V_j \text{ has a directed path to } V_i\}$ and $\mathrm{Desc}(V_i) := \{V_j : V_i \text{ has a directed path to } V_j\}$, respectively, to denote the ancestors and descendants of V_i. For DAGs, notice that $\mathrm{Anc}(V_i) \cap \mathrm{Desc}(V_i) = \{V_i\}$, but $V_i \notin \mathrm{Pa}(V_i) \cup \mathrm{Ch}(V_i)$.

We assume that only some subset $\mathcal{X} \subseteq \mathcal{V}$ is observed, with the remaining $\mathcal{L} = \mathcal{V} - \mathcal{X}$ being latent. We use V_i to denote a generic variable, observed or latent, while $X_i \in \mathcal{X}$ denotes an observed variable and $L_j \in \mathcal{L}$ denotes a latent variable. When it is clear from context, we occasionally suppress the distinction between a variable V_i and its index i.

2.2 Identification and minimality

Since $\mathbf{F}$ induces a DAG, we can always solve (1) to express the causal variables in terms of the independent noise terms:

$$\mathbf{V} = \mathbf{M}\varepsilon, \tag{2}$$

where

$$\mathbf{M} := (\mathbf{I} - \mathbf{F})^{-1} \tag{3}$$

is the mixing matrix with $\mathbf{M}_j^i$ being the net effect of ε_i on V_j. This net effect is calculated by multiplying causal weights along paths and summing across paths. Notice that if $\mathbf{M}_j^i \neq 0$, then $V_j \in \mathrm{Desc}(V_i)$.

Because $\mathbf{L}$ is hidden, let us explicitly write $\mathbf{X}$ in terms of ε:

$$\mathbf{X} = \mathbf{M}_{\mathcal{X}}\varepsilon. \tag{4}$$

In both the non-Gaussian and heterogeneous settings we consider in this paper, $\mathbf{M}_{\mathcal{X}}$ is identifiable up to permutation and scaling of columns; that is, we can identify the equivalence class

$$\mathcal{M} = \{\mathbf{M}_{\mathcal{X}}\mathbf{D}\mathbf{P} : \mathbf{D}\mathbf{P} \in \mathcal{DP}_p\}, \tag{5}$$

where

$$\mathcal{DP}_p := \{\mathbf{D}\mathbf{P} \in \mathbb{R}^{p\times p} : \mathbf{D} \text{ is full rank diagonal and } \mathbf{P} \text{ is a permutation matrix}\}.$$

We argue this for both settings individually in Section 3.

We say that an adjacency matrix $\mathbf{F}$ **generates** $\mathcal{M}$ if $(\mathbf{I} - \mathbf{F})^{-1}_{\mathcal{X}} \in \mathcal{M}$. Of course, in partially observed systems, the adjacency matrix that generates $\mathcal{M}$ is not unique. However, some of these matrices are sparser than others. In causal discovery, as in model selection more broadly, we tend to prefer the "simplest" model that adequately fits the data [4, 5]. As a result, without prior knowledge, a partially observed linear causal model cannot be identified if the population distribution can be written in terms of an equally sparse or sparser model; after all, we would never select a complicated model if a simpler model fits just as well. It is therefore natural to recast the question of identifiability to a question of maximal sparsity.

Let the ℓ_0 "norm" of a matrix $\|\cdot\|_0$ denote the number of non-zero entries in that matrix. Then we say that a causal adjacency matrix $\mathbf{F}$ is **minimal** with respect to $\mathcal{M}$ if $\mathbf{F}$ generates $\mathcal{M}$ and $\|\hat{\mathbf{F}}\|_0 \geq \|\mathbf{F}\|_0$ for any $\hat{\mathbf{F}} \neq \mathbf{F}$ that generates $\mathcal{M}$.

Let $\mathcal{F}$ denote the class of minimal adjacency matrices that generate $\mathcal{M}$. Clearly, since $\mathcal{M}$ is identifiable, so is $\mathcal{F}$. We say that an adjacency matrix $\mathbf{F}$ is **identified up to trivialities** if

$$\mathcal{F} = \left\{(\mathbf{D}\mathbf{P})^{-1}\mathbf{F}\mathbf{D}\mathbf{P} : \mathbf{D}\mathbf{P} \in \mathcal{DP}_p \text{ with } (\mathbf{D}\mathbf{P})^{\mathcal{X}}_{\mathcal{X}} = \mathbf{I}\right\}. \tag{6}$$

The only indeterminacy remaining in $\mathcal{F}$ amounts to re-indexing and re-scaling the latent factors.

A word of caution is in order. Because the adjacency matrix that generates $\mathcal{M}$ is not unique in the partially observed case, it is only possible to talk about identification with respect to some selection principle. Throughout this work we use minimality as such a selection principle – indeed we define identification in terms of it. As justification, in Section 5.1, we describe one class of non-minimal adjacency matrices which are pathological and whose exclusion is desirable; further, in Section 6, we show that existing works make assumptions even stronger than minimality; further still, in Section 7, we show that popular model selection criteria like BIC favor minimal graphs. Nevertheless, just as BIC is not always the most sensible criterion for model selection, so minimality is not always the most sensible principle for an identification theory. For example, Figure 1 shows a non-minimal graph which is not pathological. Thus, if a practitioner believes the true partially observed causal model to be non-minimal, they should content themselves with partial identification (c.f. [6]).

In Sections 4 and 5, we express identification up to trivialities in terms of two local graphical conditions, which are much easier to check than (6). But first, we return to the identifiability of $\mathcal{M}$.

3 Sufficient conditions for identification of $\mathcal{M}$

The main results of Sections 4 and 5 rely on the identifiability of $\mathcal{M}$, which is theoretically guaranteed in the two settings we consider in this paper. In the first setting, ε_i are assumed to be independent and non-Gaussian. Then according to Theorem 3 by Eriksson and Koivunen [7], $\mathcal{M}$ is identifiable from the distribution of $\mathbf{X}$. The task of estimating $\mathcal{M}$ from $\mathbf{X}$ is known as Overcomplete Independent Component Analysis (OICA) [3], and in practice this task is known to be computationally difficult [8].

In the second setting, ε_i are uncorrelated from each other with changing variances across multiple domains (or over time) and $\mathbf{M}_{\mathcal{X}}$ has full row rank (which is always the case for acyclic models). Note that in this setting, while the components of ε are mutually independent within each domain, they are not necessarily mutually independent across domains because their variances may be dependent across domains. This setting is expected to apply to a number of nonstationary scenarios including brain signal analysis, and the following theorem establishes the corresponding identifibiality of $\mathcal{M}$. Besides complementing Theorem 3 of Eriksson and Koivunen [7] as an alternative foundation for our identification work, the identifiability of $\mathcal{M}$ in this setting may be of independent interest in the fields of blind source separation and system identification.

Theorem 1. *Suppose we have observed* $\mathbf{X}$ *generated according to the mixing procedure (4) in a number of domains,* $t = 1, 2, ..., T$*, where* $\mathbf{M}_{\mathcal{X}}$ *has full row rank. Assume that* ε_i *are uncorrelated in each domain and that their variances in domain* t*, denoted by* σ_{ti}^2*, change independently across domains in the sense that* $\mathbf{S}$*, whose* (t, i)*-th entry is* σ_{ti}^2*, has full column rank. Further assume that each* $|\mathcal{X}|$ *columns of* $\mathbf{M}_{\mathcal{X}}$ *are linearly independent and that* $p \leq 2|\mathcal{X}| - 2$*. Then if* $\mathbf{X}$ *admits a model* $\mathbf{X} = \tilde{\mathbf{M}}_{\mathcal{X}}\tilde{\varepsilon}$*, where* $\tilde{\varepsilon}$ *also follows the above assumption on* ε*, then every column of* $\tilde{\mathbf{M}}_{\mathcal{X}}$ *must be proportional to a column of* $\mathbf{M}_{\mathcal{X}}$ *and vice versa.*

Note that this theorem gives sufficient conditions for the identifiability of $\mathcal{M}$; our empirical results suggest that they are not necessary.

4 Necessary conditions for identification of $\mathbf{F}$ up to trivialities

In this section, we introduce our identification conditions and show that they are necessary for $\mathbf{F}$ to be identified up to trivialities. The identification conditions are graphical conditions described in terms of "bottlenecks" and "redundancies."

Let J, K, and B be subsets of the nodes of a directed graph. Note that they need not be mutually disjoint. We say that B is a **bottleneck** from J to K if, for every $j \in J$ and every $k \in K$, each directed path from j to k includes some $b \in B$. A bottleneck B from J to K will be called **minimal** if every bottleneck B' from J to K has $|B'| \geq |B|$, and **unique minimal** if the inequality is strict for $B' \neq B$. Note that bottlenecks do not in general d-separate J and K along all paths, but only directed paths from J to K.

It is clear from the definition that, for each V_i, $\mathrm{Ch}(V_i)$ is a bottleneck from $\mathrm{Ch}(V_i)$ to X. However, for identification, we further require:

Condition 1 (Bottleneck)**.** *For every* V_i*,* $\mathrm{Ch}(V_i)$ *is the unique minimal bottleneck from* $\mathrm{Ch}(V_i)$ *to* $\mathcal{X}$*.*

As illustrated in Figure 1, the bottleneck condition ensures that if we try to "explain" the net effect of V_i on $\mathcal{X}$ by replacing $\mathrm{Ch}(V_i)$ with any subset of $\mathrm{Desc}(V_i)$, the result is a denser graph. As illustrated in Figure 3, the strong non-redundancy condition will further ensure that we cannot "explain" the effect of V_i on $\mathrm{Ch}(V_i)$ via some of its *non*-descendants:

Condition 2 (Strong Non-Redundancy)**.** *For all* L_i, V_j*, if* $\mathrm{Ch}(L_i) \subseteq \mathrm{Ch}(V_j) \cup \{V_j\}$ *then* $L_i = V_j$*.*

Figure 2 shows a graph that satisfies both of these conditions. To build intuition, let us list some simple consequences of these conditions. By the bottleneck condition, each variable must have fewer than $|\mathcal{X}|$ children; but if a variable has no latent children, then the bottleneck condition is satisfied trivially for that variable. By strong non-redundancy, each latent variable must have at least two children. For any pair (L_i, V_j), if L_i is an ancestor but not a parent of V_j, or has more than one directed path to V_j, then strong non-redundancy is satisfied for that pair. If V_j is a parent of L_i and they violate strong non-redundancy, then the bottleneck condition is violated for V_j.

Theorem 2. *If* $\mathbf{F}$ *is identified up to trivialities, then the graph induced by* $\mathbf{F}$ *satisfies the bottleneck and strong non-redundancy conditions.*

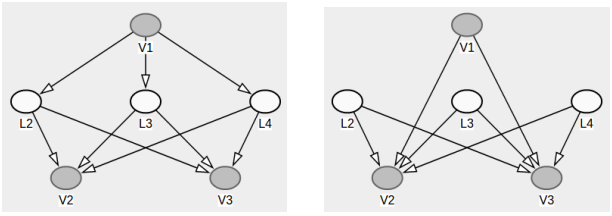


Figure 1: An egregious violation of the bottleneck condition. Left: $\{V_2, V_3\}$ is a strictly smaller bottleneck from $\mathrm{Ch}(V_1)$ to $\mathcal{X}$. Right: a sparser yet observationally equivalent graph. Although both graphs also violate strong non-redundancy, egregious bottleneck violations are not always redundant.

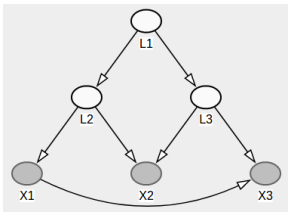
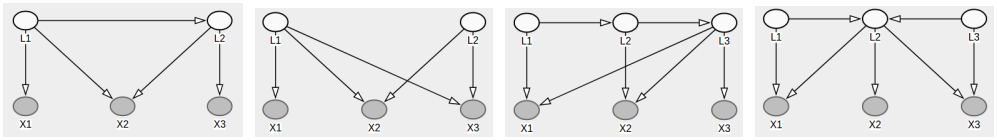


Figure 2: A simple graph illustrating our structural conditions. L_1 satisfies the bottleneck condition. L_2 and L_3 are non-redundant as each has a child the other does not. X_1 and L_3 are non-redundant as $L_3 \to X_2$ and $X_1 \notin \mathcal{L}$.

Thus the bottleneck and strong non-redundancy conditions are necessary for identification of $\mathbf{F}$ up to trivialities. In Section 5, we further show that they (along with a very mild constraint on the causal weights) are also jointly sufficient conditions.

If $\mathrm{Ch}(V_i)$ is not at least a *minimal* bottleneck for every V_i, then $\mathbf{F} \notin \mathcal{F}$. Figure 1 shows one example of such a violation of the bottleneck condition. Otherwise, as long as bottleneck faithfulness is also satisfied, $\mathcal{F}$ is an equivalence class of equally sparse latent structures which all violate at least one of the bottleneck and strong non-redundancy conditions. The nature of these indeterminacies is depicted in Figure 3. In Figure 2 we show a simple yet illustrative example in which both conditions are satisfied.



(a) (b) (c) (d)

Figure 3: Two equivalence classes. (a) and (b) are equivalent, the former violating the bottleneck condition ($\mathcal{X} \neq \mathrm{Ch}(L_1)$ is a minimal bottleneck from $\mathrm{Ch}(L_1)$ to $\mathcal{X}$) and the latter strong non-redundancy ($\mathrm{Ch}(L_2) \subseteq \mathrm{Ch}(L_1)$). (c) and (d) are equivalent, both violating strong non-redundancy.

5 Sufficient conditions for identification of $\mathbf{F}$ up to trivialities

In the previous section, we introduced two structural conditions which must be satisfied for $\mathbf{F}$ to be identifiable up to trivialities. In this section, we prove that they (along with "bottleneck faithfulness," a very mild constraint on the causal weights) are also jointly sufficient. Throughout, we assume that $\mathbf{X}$ is generated according to (1). In particular, we assume that $\mathcal{M}$ is identifiable, for example due to Theorem 3 of Eriksson and Koivunen [7] or Theorem 1 of the present work.

5.1 Bottleneck faithfulness

First, we connect ranks of submatrices of $\mathbf{M}$ to minimal bottlenecks of its corresponding graph.

Proposition 1. *Let B be a minimal bottleneck from J to K. Then* $\mathrm{Rank}(\mathbf{M}_K^J) \leq |B|$.

Strict inequality in Proposition 1 for some minimal bottleneck B from J to K can make $\mathbf{F}$ non-identifiable – even if the bottleneck condition and strong non-redundancy hold. For instance, both graphical conditions hold for

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & -1 \\ 2 & 2 & 0 \\ 3 & 3 & 0 \\ 4 & 0 & 4 \end{bmatrix} \mathbf{L} + \varepsilon_X, \quad \mathbf{L} = \varepsilon_L,$$

but $\mathrm{Rank}(\mathbf{M}_{\mathcal{X}}^{\mathcal{L}}) = 2$ while the minimal bottleneck from $\mathcal{L}$ to $\mathcal{X}$ is $\mathcal{L}$ with $|\mathcal{L}| = 3$. The system

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & -1 \\ 0 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} \mathbf{L} + \varepsilon_X, \quad \mathbf{L} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{L} + \varepsilon_L$$

generates the same mixing matrix, $\mathbf{M}_{\mathcal{X}}$, but has a strictly sparser graph. Thus to ensure identifiability, we assume that the causal coefficients satisfy:

Condition 3 (Bottleneck Faithfulness)**.** *For every $J \subseteq \mathcal{V}, K \subseteq \mathcal{X}$, if B is a minimal bottleneck from J to K, then* $\mathrm{Rank}\left(\mathbf{M}_K^J\right) = |B|$.

In the supplementary material we characterize the set of adjacency matrices $\mathbf{F}$ that are bottleneck faithful for a given graph. In particular, we show that a generic $\mathbf{F}$ is bottleneck faithful.

Interestingly, in linear systems, classical faithfulness is a special case of bottleneck faithfulness. $\mathrm{Rank}(\mathbf{M}_K^J) = 0$ is a violation of classical faithfulness if there is a minimal bottleneck $B \neq \varnothing$ from

J to K. That is, if there is a path from J to K but the path coefficients cancel out so that the net effect of J on K is 0, the system is not faithful to the graph. Bottleneck faithfulness generalizes this so that the net effect of J on K must have maximal rank for the given graph.

5.2 Identifiability

In this subsection, we show that if the bottleneck condition, strong non-redundancy, and bottleneck faithfulness hold for $\mathbf{F}$, then $\mathbf{F}$ is identifiable up to trivialities. Throughout, we assume the three conditions hold.

Our approach is illustrated by the following computation. For any V_i,

$$\mathbf{M}_{\mathcal{X}}(\mathbf{I}-\mathbf{F})^i = \mathbf{I}^i_{\mathcal{X}}. \tag{7}$$

Let $J = \mathrm{Ch}(V_i)$. Since the support of $\mathbf{F}^i$ is J, the equation

$$(\mathbf{M}-\mathbf{I})^i_{\mathcal{X}} = \mathbf{M}^J_{\mathcal{X}}\mathbf{x}. \tag{8}$$

always has a solution at $\mathbf{x} = \mathbf{F}^i_J$. In fact, (under the three assumptions of this section) this solution is unique:

Lemma 1. *Let $J = \mathrm{Ch}(V_i)$. Then the unique solution to* (8) *is given by $\mathbf{x} = \mathbf{F}^i_{\mathrm{Ch}(V_i)}$.*

But there is a version of (8) for each $J \subseteq \mathcal{V}$. For which other choices of J does (8) have a solution? Clearly a solution always exists if $J \supseteq \mathrm{Ch}(V_i)$. On the other hand, we can guarantee that a solution with $|J| \leq |\mathrm{Ch}(V_i)|$ is only possible if J contains an ancestor of V_i:

Lemma 2. *Suppose $J \subseteq \mathcal{V} - \mathrm{Anc}(V_i)$. If $(\mathbf{M}-\mathbf{I})^i_{\mathcal{X}} \in \mathrm{Range}\left(\mathbf{M}^J_{\mathcal{X}}\right)$, then $|J| \geq |\mathrm{Ch}(V_i)|$, with equality if and only if $J = \mathrm{Ch}(V_i)$.*

By Lemma 2, if any superset of $\mathrm{Ch}(V_i)$ containing no ancestors of V_i is identifiable, then $\mathrm{Ch}(V_i)$ is also identifiable. Next, we will show how such a superset of $\mathrm{Ch}(V_i)$ can be identified.

Let $\mathcal{V}_k \subseteq \mathcal{V}$ denote the variables whose longest path to $\mathcal{X}$ has fewer than k nodes. More formally, we define recursively

$$\mathcal{V}_0 := \varnothing, \tag{9}$$

$$\mathcal{V}_{k+1} := \{V_i \in \mathcal{V} : \mathrm{Ch}(V_i) \subseteq \mathcal{V}_k\}, \quad \text{for } k \geq 0. \tag{10}$$

Naturally, we define $\mathcal{X}_k := \mathcal{V}_k \cap \mathcal{X}$ and $\mathcal{L}_k := \mathcal{V}_k \cap \mathcal{L}$. Notice that $\mathcal{V}_k$ is strictly increasing, and is induced by the topological ordering on $\mathcal{V}$.

Proposition 2. *For all k, either $\mathcal{V}_k \subset \mathcal{V}_{k+1}$, or $\mathcal{V}_k = \mathcal{V}$.*

Further, for each $k \geq 0$, define

$$\mathcal{J}_{k+1}(V_i) := \underset{J \in \left\{J \subseteq \mathcal{V}_k : \mathbf{M}^i_{\mathcal{X}_k} \in \mathrm{Range}\left(\mathbf{M}^J_{\mathcal{X}_k}\right)\right\}}{\arg\min} |J|. \tag{11}$$

Intuitively, this denotes the set of minimal choices for $J \subset \mathcal{V}_k$ such that (8) has a solution. From Lemma 2, we know that $\mathcal{J}_{k+1}(V_i) = \{\mathrm{Ch}(V_i)\}$ if $V_i \in \mathcal{V}_{k+1} - \mathcal{V}_k$. The construction of (11) allows us to generalize Lemma 2 to describe which versions of (8) have solutions when we are not sure of the causal order.

Lemma 3. *For every $k \geq 0$, let $\mathcal{V}_k$ and $\mathcal{J}_k(V_i)$ be defined as above. Then $V_i \in \mathcal{V}_{k+1} - \mathcal{V}_k$ if and only if all of the following hold:*

1. *$V_i \notin \mathcal{V}_k$;*
2. *$|\mathrm{Support}(\mathbf{M}^i_{\mathcal{X}}) - \mathcal{X}_k| \leq 1$;*
3. *$|\mathcal{J}_{k+1}(V_i)| = 1$; and*
4. *for all $V_j \neq V_i$ satisfying points 1 and 2, $\mathbf{M}^j_{\mathcal{X}_k} \notin \mathrm{Range}\left(\mathbf{M}^{J_k(V_i)}_{\mathcal{X}_k}\right)$.*

As a result of Lemma 3, each $\mathcal{V}_k$ is identifiable. Clearly, if $V_i \in \mathcal{V}_{k+1}$, then $\mathrm{Ch}(V_i) \in \mathcal{V}_k$ and $\mathcal{V}_k \cap \mathrm{Anc}(V_i) = \varnothing$. Hence by Lemma 2, $\mathrm{Ch}(V_i)$ is also identifiable. Thus the full DAG is identifiable, and each column of $\mathbf{M}_{\mathcal{X}}$ can be associated with the corresponding node in the DAG. However, Lemmas 1, 2, and 3 are not enough to distinguish which nodes correspond to latent variables and which correspond to observed variables; we have yet to pair each X_i with its net effects $\mathbf{M}^i_{\mathcal{X}}$. Resolving this final indeterminacy is not hard. Intuitively, the vector of $\mathbf{M}_{\mathcal{X}}$ corresponding to X_i must have non-zero coefficients in the i-th slot while every vector corresponding to descendants of X_i will not. Lemma 4 formalizes this observation.

Lemma 4. *$X_i \in \mathcal{X}_{k+1}$ if and only if $X_i \in \mathcal{V}_{k+1}$ and* $\mathrm{Support}\,(\mathbf{M}^{X_i}_{\mathcal{X}}) - \mathcal{X}_k = \{i\}$.

Together, Lemmas 1, 2, 3, and 4 imply that $\mathbf{F}$ is identifiable if $\mathbf{M}_{\mathcal{X}}$ is identifiable. Of course, we do not know $\mathbf{M}_{\mathcal{X}}$—only $\mathcal{M}$. Nevertheless, Lemmas 2, 3, and 4 do not involve the scaling and permutation of $\mathbf{M}_{\mathcal{X}}$—only the linear dependencies of its columns. Some simple calculation shows that Lemma 1 can be used to put any $\mathbf{M}_{\mathcal{X}}\mathbf{PD} \in \mathcal{M}$ in one-to-one correspondence with $(\mathbf{PD})^{-1}\mathbf{F}(\mathbf{PD})$.

Theorem 3. *Suppose $\mathbf{F}$ satisfies generalized non-redundancy, bottleneck faithfulness, and the bottleneck condition. Then $\mathbf{F}$ is identifiable up to indeterminacies.*

6 Relation to existing work

Constraint- and score-based approaches to causal discovery based on conditional independence testing—such as SGS [1], IC [2], PC [9], GES [10], and FGS [11]—generally focus on the causally sufficient case. These algorithms identify the Markov equivalence class of graphs which all encode the same set of conditional independence relations. While some methods based on conditional independence tests, such as FCI [12] and RFCI [13], are able to relax the assumption of causal sufficiency, their focus is on learning the causal relations between observed variables and distinguishing them from spurious dependencies induced by shared latent ancestors. Such methods recover only limited information about the latent structure, as only the most basic information about latent structure is identifiable from conditional independence relations alone. For one review of causal discovery methods, see Spirtes and Zhang [14].

It is possible to go beyond the equivalence class with additional assumptions on causal mechanisms [14]. In particular, linear non-Gaussian models have been studied extensively. In the causally sufficient case, Shimizu et al. [15] leverages acyclicity of the causal relations and the identifiability of the square ICA problem [16, 17] to show how the causal adjacency matrix can be identified, while Lacerda et al. [18] further estimate a subclass of cyclic causal models. In both cases, one may replace the non-Gaussian noise assumption with the heterogeneous noise assumption (in the formal sense of Theorem 1) and the identifiability results still hold [19].

By contrast, previous works on partially observed linear non-Gaussian models only study certain special cases in which the models are partially identifiable. Hoyer et al. [20] describe a procedure to convert partially observed causal models to a canonical form in which no latent variable has any parents. They further provide an algorithm which recovers all canonical forms consistent with the observed overcomplete basis $\mathcal{M}$, which is identifiable by OICA [3]. This recovered equivalence class of observationally equivalent canonical forms can be huge, and by definition can neither identify causal relations among latent confounders nor distinguish latent confounders from latent mediators.

More recently, Lemma 5 of Salehkaleybar et al. [6] states that if $\mathcal{M}$ is identifiable, then the causal order among observed variables is identifiable if classical faithfulness holds between all variables. Their condition is strictly weaker than ours; as we discuss in Section 5.1, classical faithfulness is entailed by bottleneck faithfulness and imposes no graphical conditions. That a weaker condition suffices for their task is not surprising, since their task is strictly easier than ours; if $\mathbf{F}$ is identified up to trivialities then the causal order among observed variables is also identified (while the causal order alone tells very little about $\mathbf{F}$). Lemma 5 has a reassuring consequence for our work: even if the graphical conditions for total identification fail to apply, the causal order of $\mathcal{X}$ is still identified.

If the practitioner is further interested in the observed variables' net effects on one another, $(\mathbf{M} - \mathbf{I})^{\mathcal{X}}_{\mathcal{X}}$, then additional graphical assumptions are needed. Theorem 16 of Salehkaleybar et al. [6] provides one condition sufficient for this purpose: no latent variable L_i has precisely the same observed descendants as any observed variable X_j (formally, for all $L_i, X_j, \mathrm{Desc}(L_i) \cap \mathcal{X} \neq \mathrm{Desc}(X_j) \cap \mathcal{X}$).

Again, this being a relatively easy subtask of the problem we consider, it is not surprising that our conditions are not strictly weaker. With that said, our conditions are not strictly stronger, either; the bottleneck and strong non-redundancy conditions can be satisfied even when L_i and X_j have the same observed descendants, as shown in Figure 4.

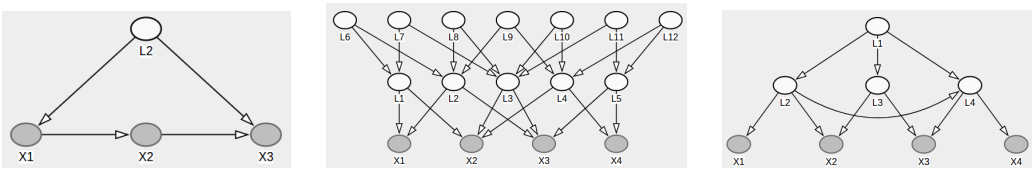


Figure 4: Examples of graphs identifiable from $\mathcal{M}$. From left to right: a graph where $\mathrm{Desc}(L) \cap \mathcal{X} = \mathrm{Desc}(X_1) \cap \mathcal{X}$; a widening hierarchical structure; a hierarchical structure with intra-layer relations.

To recover causal structures of the hidden variables, many results rely on strong assumptions about clusters of pure variables (sets of observed variables which each share a latent confounder and have no other parents). For example, Spearman's classical Tetrad condition [21] identifies latent causes with four pure observed children from covariance information alone. In the linear non-Gaussian case, existing work reduces the number of pure observed children to three [22], and more recently to two [23, 24]. Clearly, these are all special cases of both the bottleneck and non-redundancy conditions. As such, our graphical assumptions are strictly weaker.

Proposition 3. *Suppose each L_i in a partially observed DAG has at least two pure children (latent or observed). Then the DAG satisfies the bottleneck and non-redundancy conditions.*

However, identification is possible even when no latent confounder has any pure children; for example, Anandkumar et al. [25] present a model in which latent variables with no pure children are identifiable. Rather than purity, they require a graph expansion property—for all non-singleton $S \subseteq \mathcal{L}$, $|\bigcup_{L_i \in S} \mathrm{Ch}(L_i) \cap \mathcal{X}| \geq |S| + d_{\max}$, where $d_{\max} = \max_i |\mathrm{Ch}(L_i) \cap \mathcal{X}|$ —as well as a rank condition on $\mathbf{F}^{\mathcal{L}}_{\mathcal{X}}$ which places hard-to-check graphical constraints on the model and bounds $|\mathcal{L}| \leq \frac{1}{3}|\mathcal{X}|$. Non-redundancy among latent variables can be derived from the expansion property by considering the case where $|S| = 2$, and the bottleneck condition by considering $S = \{L_i\} \cup \mathrm{Ch}(L_i) \cap \mathcal{L}$. Thus in one sense, our conditions can be seen as a refinement on the expansion property; however, we also remove the many hard-to-check graphical consequences of the rank condition, and further show that many graphs even with $|\mathcal{L}| \gg |\mathcal{X}|$ are identifiable. For example, Figure 4 shows an identifiable hierarchical model in which the number of latent variables increases with depth. Moreover, we show that our conditions are sufficient for identifying hierarchical structures in which variables in the same layer are causally related. Figure 4 shows one such system.

Proposition 4. *Suppose* $\mathbf{F}$ *satisfies the rank and graph expansion conditions of Anandkumar et al. [25]. Then* $\mathbf{F}$ *also satisfies the bottleneck and strong non-redundancy conditions.*

Propositions 3 and 4 show that our conditions are indeed more general than previous identification conditions; not only do our conditions allow and identify causal relations among observed variables, they also identify latent structures which no previous works could. (See, for example, Figure 4.) Furthermore, in light of Theorems 2 and 3, they show that many existing works implicitly took sparsity of causal edges as a useful primitive for what it means for a partially observed causal model to be identifiable. Such a primitive is widely used throughout causal discovery, even in the causally sufficient case [4, 5].

Although many of these conditions for latent structure identification rely on non-Gaussian independent noise, direct estimation of the mixing matrix is often avoided in practice, especially in the causally insufficient case, as estimation of the overcomplete mixing matrix is computationally challenging [8]. Estimation of the mixing matrix can be avoided by directly using the independent additive noise assumption and exploiting graphical conditions such as causal sufficiency or purity. For example, in the causally sufficient case, $\mathrm{Pa}(X_i)$ is identifiable by regressing X_i on $\mathcal{Z} \subseteq \mathcal{X}$ and testing the independence of the regression residuals and $\mathcal{Z}$ [26]. This approach may be adapted to the non-linear [27, 28] and post-nonlinear [29] cases. Tashiro et al. [30] extend this idea to identify causal relations among observed variables in the causally insufficient case, and a related condition is developed by Xie et al. [24] to identify one special type of confounder. However, such methods owe their efficiency to the strong structural conditions under which they guarantee identifiability. As the bottleneck

and non-redundancy conditions are much more general, this naturally complicates the question of estimation.

7 Estimation

In Theorems 1 and 3, we have shown that a causal system which satisfies the conditions of Section 4 is uniquely identifiable whenever $\mathbf{M}_{\mathcal{X}}$ is identifiable. However, as indicated in Section 6, estimation of $\mathbf{M}_{\mathcal{X}}$ from homogeneous non-Gaussian data—for example, by overcomplete ICA—is computationally hard. Further, whereas the estimation algorithms presented in [15], [18], and [6] require the practitioner to test which entries of $\mathbf{M}_{\mathcal{X}}$ are exact zeros, a naive algorithm inspired by Lemmas 1, 2, 3, and 4 would further require them to test which submatrices' singular values are exact zeros. Such an algorithm is not advisable.

As a proof of concept, we therefore focus our experiments on partially observed linear causal models in the heterogeneous case. In this setting, $\mathbf{F}$ can be learned directly by optimizing the regularized likelihood with respect to $\mathbf{F}$, given the sample covariance matrices of $\mathbf{X}$. We leave more efficient estimation in more general settings to future work.

7.1 Simulations

Suppose we have access to samples from T heterogeneous domains. The data in the t-th domain follow

$$\mathbf{V} = \mathbf{F}\mathbf{V} + \varepsilon, \tag{12}$$

where $\varepsilon \sim \mathcal{N}(0, \Sigma_t)$ for diagonal Σ_t. Then in the t-th domain,

$$\mathbf{X} = \mathbf{M}_{\mathcal{X}} \Sigma_t \mathbf{M}_{\mathcal{X}}^T. \tag{13}$$

The negative log likelihood is

$$-2\ell\ell(\mathbf{F}, \Sigma) = \sum_{t=1}^{T} n_t \left(|\mathcal{X}| \log(2\pi) + \log\det(\mathbf{S}_t) + \operatorname{Tr}\left(\mathbf{S}_t^{-1} \hat{\mathbb{E}}\left[\mathbf{x}_t \mathbf{x}_t^T \right] \right) \right), \tag{14}$$

where $\mathbf{x}_{t,i}$ is the i-th row of the design matrix for the t-th domain, $\hat{\mathbb{E}}\left[\mathbf{x}_t \mathbf{x}_t^T\right]$ is the empirical second moment of the d-th domain, $\mathbf{S}_t = \mathbf{M}_{\mathcal{X}} \Sigma_t \mathbf{M}_{\mathcal{X}}^T$, and n_t is the sample size in the t-th domain. If $\mathbf{X}$ is generated according to (12), the independent change condition in Theorem 1 holds for the noise variances, and $\mathbf{F}$ satisfies the assumptions of Section 4, then by Theorems 1 and 3, $\mathbf{F}$ is identifiable up to trivialities. Hence we can in principle optimize the regularized log likelihood.

As a sanity check for our theoretical results, we simulate data according to (12); for every identifiable graph structure with three observed variables and at most five directed edges, we generated ten causal adjacency matrices with weights randomly drawn from $(-0.9, -0.5) \cup (0.5, 0.9)$. We estimate $\mathbf{F}$ and $|\mathcal{L}|$ by minimizing the BIC via exhaustive search. By enumerating candidate graphs from sparsest to densest, a single search could take anywhere from 10 to 60 minutes on an Intel core i7 processor.

To verify that our estimation method was actually leveraging the noise's heterogeneity, we ran the experiment with one domain and 5000 observations. Only 3% of graphs were identified. Not surprisingly, only 15% of learned graphs had any latent variables at all. Increasing the number of domains from 1 to 3 but keeping the total sample size at 5000 (i.e. 1666 per domain) improved the rate of structure identification to 50% of trials.

With 5 domains and 500 samples per domain, the correct graph is identified on 50% of trials; with 1000 samples per domain, this improves to 70%; and with 10 000, this further improves to 80%. In every case that the wrong graph was recovered, the equivalence class of mixing matrices $\hat{\mathcal{M}}$ generated by $\hat{\mathbf{F}}$ had incorrect support, perhaps due to insufficient domains or accidentally coupled changes in the noise variances. This supports the main theory of Theorem 3, which, in light of Theorem 1, claims that the structure of $\mathbf{F}$ is uniquely determined *from a correctly identified* $\mathcal{M}$. We report detailed results in the supplement for all graphs studied.

To verify our claims in Theorem 2, we also tested simulated data from ten non-identifiable partially observed DAGs—six of which stand in the main equivalence class relations of Figure 3, and

four of which are not minimal. Not surprisingly, members of the same equivalence class were indistinguishable, each system achieving the same log likelihood up to eight significant digits.

Because exhaustive search over graphs is computationally expensive (even for this toy problem, there are 1759 graphs, and so 1759 non-convex optimization problems), it would be desirable to instead optimize the L1-penalized negative log likelihood:

$$\mathrm{L}(\mathbf{F}, \Sigma) = -2\ell\ell(\mathbf{F}, \Sigma) + \lambda \sum |F_{i,j}|. \tag{15}$$

As before, we simulated data from (12). Numerical experiments verify that $\mathrm{L}(\mathbf{F}^0, \Sigma^0)$ is very near a local minimum for all practical $\lambda > 0$, where $(\mathbf{F}^0, \Sigma^0)$ denotes the true adjacency matrix and noise covariances. However, experiments also suggest that this local minimum is generally quite far from the global minimum, both in parameter space and in L1 loss. Moreover, while the L1 penalty successfully drives many parameters to zero (as we would expect), our experiments frequently converge to minima which are denser than the true system. Intuitively, the L1 penalty does not care about the density of $\hat{\mathbf{F}}$; a dense system with small coefficients may have a comparable L1 penalty as a sparse system with large coefficients. Naturally, denser systems are better equipped to fit the observed domain covariances. We summarize these experiments in the supplement.

8 Conclusion and discussions

In many fields, we do not believe that all causally relevant variables have been measured. In such partially observed settings, beyond accurately estimating the causal relations among observed variables, practitioners may want to further identify the causal relations among the hidden variables which generate the observed data. Inspired by this issue, we have contributed to the identification theory of partially observed linear causal models by providing necessary and sufficient graphical conditions for the identification of the full causal graph. Throughout, we assume the additive noise terms in the structural equation model follow non-Gaussian distributions or have independently changing variances across time or between domains. Such assumptions, unlike the single-domain Gaussianity assumption, render the mixing procedure from the noise terms to the observed variables identifiable up to the permutation and scaling of columns, thereby facilitating our final identifiability results. These conditions are expected to be applicable to a wide variety of partially observed structures. To deal with real applications, efficient estimation methods are needed, and we hope our theoretical identifiability results will stimulate algorithmic development to finally solve this important causal discovery problem. As future work, we will focus on developing practical estimation methods and extending our results to nonlinear cases.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

The work presented in this article was supported in part by Novo Nordisk Foundation Grant NNF20OC0062897.

This work was supported in part by the National Institutes of Health (NIH) under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award #2134901, by the United States Air Force under Contract No. FA8650-17-C7715, and by a grant from Apple Inc. The NIH or NSF is not responsible for the views reported in this article.

We are grateful to the anonymous reviewers for their careful reading and helpful comments.

References

[1] Peter Spirtes, Clark Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

[2] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[3] Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[4] Malcolm Forster, G. Raskutti, Reuben Stern, and Naftali Weinberger. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 69, 04 2017.

[5] Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graphs based on sparsest permutations. *Stat*, 2018.

[6] Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-Gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

[7] J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

[8] Chenwei Ding, Mingming Gong, Kun Zhang, and Dacheng Tao. Likelihood-free overcomplete ICA and applicationsin causal discovery. In *Advances in Neural Information Processing Systems*, 2019.

[9] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[10] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[11] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

[12] Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, Causation, and Discovery*, 1999.

[13] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[14] Peter Spirtes and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 2016.

[15] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 2006.

[16] P. Comon. Independent Component Analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[17] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

[18] G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by Independent Components Analysis. *Uncertainty in Artificial Intelligence*, 2008.

[19] Kiyotoshi Matsuoka, Masahiro Ohoya, and Mitsuru Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[20] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[21] Charles Spearman. Pearson's contribution to the theory of two factors. *British Journal of Psychology*, 19(1):95, 1928.

[22] Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

[23] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in Neural Information Processing Systems*, 2019.

[24] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating linear non-Gaussian latent variable graphs. *Advances in Neural Information Processing Systems*, 2020.

[25] Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham M. Kakade. Learning linear Bayesian networks with latent variables. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[26] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[27] PO. Hoyer, D. Janzing, JM. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, Red Hook, NY, USA, June 2009. Max-Planck-Gesellschaft, Curran.

[28] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

[29] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

[30] Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural Computation*, 2014.

Supplementary Material: Identification of Partially Observed Linear Causal Models

Jeffrey Adams[1], **Niels Richard Hansen**[1], **Kun Zhang**[2]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark
[2]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
ja@math.ku.dk, niels.r.hansen@math.ku.dk, kunz1@cmu.edu

1 Proofs and more details

For convenience, we reserve the matrix $\mathbf{R}_B^J$ to denote the net effect of J on B *prior* to any mixing effects among B—that is, on the subgraph where $\mathrm{Ch}(V_b)$ has been set to $\varnothing$ for each $V_b \in B$.

Proposition 3. *For any $J, B \subseteq \mathcal{V}$, $\mathbf{R}_B^J = \left(\mathbf{M}_B^B\right)^{-1} \mathbf{M}_B^J$. Moreover, if B is a bottleneck from J to K, then $\mathbf{M}_K^J = \mathbf{M}_K^B \mathbf{R}_B^J$.*

1.1 Theorem 1 and its proof

Let us present the complete theorem first, and then give its proof. Let n be the dimensionality of $\mathbf{X}$. Remeber p is the number of noise terms. In the case where $n = p$, $\mathbf{M}_{\mathcal{X}}$ in (4) is a square matrix, and its identifiability from $\mathbf{X}$ up to column rescaling and permutations has been provided by Matsuoka et al. [1], but we are concerned with the case where $n < p$. In the case where the noise terms ε_i are non-Gaussian, the identifiability of $\mathbf{M}_{\mathcal{X}}$ up to column rescaling and permutations was also given in the literature [2], inspired by the results in [3]. Although the corresponding OICA problem may be difficult to solve in practice, this identifiability result is nice in that it holds true even if p is much larger than n. The heterogeneous variance case seems complementary: its maximum likelihood estimation procedure is simple, but our proof of it uses a constraint on p, given a fixed n (this condition is sufficient, but may be unnecessary, as illustrated by our simulation results), as given in the following theorem.

Before presenting Theorem 1, let us give the following lemma, which will be needed in the proof of Theorem 1.

Lemma 5. *Suppose matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ has linearly independent columns, i.e., Rank$(\mathbf{K}) = n$. Let $\mathring{\mathbf{K}} = \mathbf{K} - d \cdot \mathbf{1}^\intercal$, where $d \in \mathbb{R}^n$ and $\mathbf{1}$ is the length-n vector of all 1's. Then for any d, Rank$(\mathring{\mathbf{K}}) \geq n - 1$.*

Proof. Since $\mathbf{K}$ in invertible, let $f := \mathbf{K}^{-1} \cdot d$. Then $\mathring{\mathbf{K}} = \mathbf{K} - d \cdot \mathbf{1}^\intercal = \mathbf{K}(\mathbf{I} - f \cdot \mathbf{1}^\intercal)$, where $\mathbf{I}$ denotes the identity matrix. Since $\mathbf{K}$ has full rank, $\mathrm{Rank}(\mathring{\mathbf{K}}) = \mathrm{Rank}(\mathbf{K}(\mathbf{I} - f \cdot \mathbf{1}^\intercal)) = \mathrm{Rank}(\mathbf{I} - f \cdot \mathbf{1}^\intercal)$.

To show $\mathrm{Rank}(\mathbf{I} - f \cdot \mathbf{1}^\intercal) \geq n - 1$, we can equivalently show that the nullspace of $(\mathbf{I} - f \cdot \mathbf{1}^\intercal)$ has at most dimension one. suppose that g is a nonzero vector in $\mathbb{R}^n$ that satisfies the equation:

$$(\mathbf{I} - f\mathbf{1}^\intercal)g = 0 \iff g = f \cdot \mathbf{1}^\intercal g,$$

which also implies $\mathbf{1}^\intercal \cdot g = \mathbf{1}^\intercal \cdot f \cdot \mathbf{1}^\intercal g$, or $\mathbf{1}^\intercal \cdot f = 1$. Therefore, there are two cases to consider:

- If the value of d satisfies $\mathbf{1}^\intercal \cdot f = \mathbf{1}^\intercal \mathbf{K}^{-1} \cdot d = 1$, the nullspace of $(\mathbf{I} - f\mathbf{1}^\intercal)$ is span(f), which has dimension one, and accordingly $\mathrm{Rank}(\mathbf{I} - f \cdot \mathbf{1}^\intercal) = \mathrm{Rank}(\mathring{\boldsymbol{K}}) = n - 1$.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

- If the value of d does not satisfy $\mathbf{1}^\intercal \mathbf{K}^{-1} \cdot d = 1$, the nullspace of $(\mathbf{I} - f\mathbf{1}^\intercal)$ has dimension zero, and consequently $\text{Rank}(\mathbf{I} - f \cdot \mathbf{1}^\intercal) = \text{Rank}(\mathring{\boldsymbol{K}}) = n$.

□

We are now ready to present Theorem 1.

Theorem 1 *Suppose we have observed* $\mathbf{X}$ *generated according to the mixing procedure (4) in a number of domains,* $t = 1, 2, ..., T$. *Assume that* ε_i *are uncorrelated in each domain and that their variances in domain* t, *denoted by* σ_{ti}^2, *change independently across domains in the sense that* $\mathbf{S}$, *whose* (i, t)*th entry is* σ_{ti}^2, *has full column rank. Further assume that each* n *columns of* $\mathbf{M}_\mathcal{X}$ *are linearly independent and that* $p \leq 2n - 2$. *Then if* $\mathbf{X}$ *admits a model*

$$\mathbf{X} = \tilde{\mathbf{M}}_\mathcal{X} \tilde{\varepsilon}, \tag{16}$$

where $\tilde{\varepsilon}$ *also follows the above assumption on* ε, *every column of* $\tilde{\mathbf{M}}_\mathcal{X}$ *must be proportional to a column of* $\mathbf{M}_\mathcal{X}$ *and vice versa.*

Proof. Let σ_{ti}^2 be the variance of $\tilde{\varepsilon}_i$ in the tth domain. Let S_t be the diagonal matrix with $\sigma_{t1}^2, \sigma_{t2}^2, ..., \sigma_{tp}^2$ on its diagonal, and similarly for $\tilde{S}_t$. Let $\tilde{\mathbf{S}}$ be the matrix with $\tilde{\sigma}_{ti}^2$ as its (i, t)th entry. Denote by $\mathbf{M}_\mathcal{X}^i$ the ith column of $\mathbf{M}_\mathcal{X}$, and similarly for $\tilde{\mathbf{M}}_\mathcal{X}^i$. In the t-th domain the two mixing models imply the same distribution, or more specifically, the same covariance matrix, of $\mathbf{X}$. That is, in the t-th domain,

$$\text{Cov}(\mathbf{X}_t) = \mathbf{M}_\mathcal{X} S_t \mathbf{M}_\mathcal{X}^\intercal = \tilde{\mathbf{M}}_\mathcal{X} \tilde{S}_t \tilde{\mathbf{M}}_\mathcal{X}^\intercal, \quad \text{or equivalently,} \tag{17}$$

$$\text{Cov}(\mathbf{X}_t) = \sum_{i=1}^{p} \sigma_{ti}^2 \mathbf{M}_\mathcal{X}^i \mathbf{M}_\mathcal{X}^{i\intercal} = \sum_{i=1}^{p} \tilde{\sigma}_{ti}^2 \tilde{\mathbf{M}}_\mathcal{X}^i \tilde{\mathbf{M}}_\mathcal{X}^{i\intercal}. \tag{18}$$

It can also be written as

$$\sum_{i=1}^{p} \sigma_{ti}^2 \mathbf{M}_\mathcal{X}^i \otimes \mathbf{M}_\mathcal{X}^i = \sum_{i=1}^{p} \tilde{\sigma}_{ti}^2 \tilde{\mathbf{M}}_\mathcal{X}^i \otimes \tilde{\mathbf{M}}_\mathcal{X}^i, \text{ or in matrix form,}$$

$$(\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}) \cdot \mathbf{S} = (\tilde{\mathbf{M}}_\mathcal{X} \odot \tilde{\mathbf{M}}_\mathcal{X}) \cdot \tilde{\mathbf{S}}, \tag{19}$$

where $\otimes$ denotes the Kronecker product and $\odot$ the Khatri–Rao (column-wise Kronecker) product, i.e., $\tilde{\mathbf{M}}_\mathcal{X} \odot \tilde{\mathbf{M}}_\mathcal{X} = [\mathbf{M}_\mathcal{X}^1 \otimes \mathbf{M}_\mathcal{X}^1,\ \mathbf{M}_\mathcal{X}^2 \otimes \mathbf{M}_{\mathcal{X}2}, ..., \mathbf{M}_\mathcal{X}^p \otimes \mathbf{M}_\mathcal{X}^p]$.

Since $\mathbf{S}$ has full column rank, we can select p columns from it, corresponding to p domains, that form a full rank matrix. Let this matrix be $\mathbf{S}^*$. Similarly we have $\tilde{\mathbf{S}}^*$ corresponding to the alternative model (16), corresponding to the same p domains. Equation (19) then implies

$$(\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}) \cdot \mathbf{S}^* = (\tilde{\mathbf{M}}_\mathcal{X} \odot \tilde{\mathbf{M}}_\mathcal{X}) \cdot \tilde{\mathbf{S}}^*, \tag{20}$$

We will use the concept Kruskal-rank [4], denoted by Rank_k; the Kruskal-rank of a matrix K is the maximum number of l such that every l columns of K are linearly independent. Bear in mind that each n columns of $\mathbf{M}_\mathcal{X}$ are linear independent (i.e., $\text{Rank}_k(\mathbf{M}_\mathcal{X}) = n$) and that $p \leq 2n - 2$. Lemma 1 by Sidiropoulos et al. [5] then implies that the rank of $\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}$ is larger than or equal to $\min(2n - 1, p) = p$. That is, $\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}$ has full column rank. Further because $\mathbf{S}^*$ has full rank, (20) implies that $\tilde{\mathbf{S}}^*$ has full rank and that $\tilde{\mathbf{M}}_\mathcal{X} \odot \tilde{\mathbf{M}}_\mathcal{X}$ has full column rank.

Right-multiplying both sides of (20) by $\tilde{\mathbf{S}}^{*-1}$ and let $\mathbf{Q} := \mathbf{S}^* \cdot \tilde{\mathbf{S}}^{*-1}$, one will get

$$(\tilde{\mathbf{M}}_\mathcal{X} \odot \tilde{\mathbf{M}}_\mathcal{X}) = (\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}) \cdot \mathbf{Q}. \tag{21}$$

We shall then show that $\mathbf{Q}$ must be a generalized permutation matrix and hence the columns of $\tilde{\mathbf{M}}_\mathcal{X}$ are a permuted and scaled version of those of $\mathbf{M}_\mathcal{X}$.

Without loss of generality, let us consider the first column of the matrices on both sides of (21), and let q_{i1} be the $(i, 1)$th entry of $\mathbf{Q}$. We have

$$\tilde{\mathbf{M}}_\mathcal{X}^1 \otimes \tilde{\mathbf{M}}_\mathcal{X}^1 = (\mathbf{M}_\mathcal{X} \odot \mathbf{M}_\mathcal{X}) \cdot \mathbf{Q}^1 = \sum_{i=1}^{p} q_{i1} \cdot (\mathbf{M}_\mathcal{X}^i \otimes \mathbf{M}_\mathcal{X}^i), \tag{22}$$

where q_{i1} cannot be zero for all i. Since each n columns of $\mathbf{M}_{\mathcal{X}}$ are linearly independent, it cannot contain a zero column. Suppose $\mathbf{M}^i_{\mathcal{X}k}$, the (k,i)th entry of $\mathbf{M}_{\mathcal{X}}$, is nonzero. According to the specific structure of the Kronecker product $\tilde{\mathbf{M}}^1_{\mathcal{X}} \otimes \tilde{\mathbf{M}}^1_{\mathcal{X}}$ in (22), we know that there must exist a non-zero vector $d \in \mathbb{R}^n$ such that the RHS satisfies

$$\begin{aligned}
&\sum_{i=1}^{p} q_{i1} \cdot (\mathbf{M}^i_{\mathcal{X}} \otimes \mathbf{M}^i_{\mathcal{X}}) = d \otimes \big(\sum_{i=1}^{p} q_{i1} \cdot (\mathbf{M}^i_{\mathcal{X}k} \cdot \mathbf{M}^i_{\mathcal{X}})\big) = \sum_{i=1}^{p} q_{i1} \cdot \big((\mathbf{M}^i_{\mathcal{X}k} \cdot d) \otimes \mathbf{M}^i_{\mathcal{X}}\big) \\
\Longrightarrow &\sum_{i=1}^{p} q_{i1} \cdot \big((\mathbf{M}^i_{\mathcal{X}} - \mathbf{M}^i_{\mathcal{X}k} \cdot d) \otimes \mathbf{M}^i_{\mathcal{X}}\big) = 0 \\
\Longrightarrow &(\mathring{\mathbf{M}}_{\mathcal{X}} \odot \mathbf{M}_{\mathcal{X}})\mathbf{Q}^1 = 0, \qquad (23)
\end{aligned}$$

where $\mathring{\mathbf{M}}_{\mathcal{X}}$ is a $n \times n$ matrix with $(\mathbf{M}^i_{\mathcal{X}} - \mathbf{M}^i_{\mathcal{X}k} \cdot d)$ as its i-th column, i.e.,

$$\mathring{\mathbf{M}}_{\mathcal{X}} = \mathbf{M}_{\mathcal{X}} - \mathbf{M}^i_{\mathcal{X}k} \cdot d \cdot \mathbf{1}^T.$$

We are now about to show that in order for (23) to hold, $q_{i1} \neq 0$ for one and only one $i = 1, 2, ..., p$. There are two cases to consider:

- Suppose one column of $\mathring{\mathbf{M}}_{\mathcal{X}}$ is zero. Note that since $\text{Rank}_k(\mathbf{M}_{\mathcal{X}}) = n$, each pair of its columns are linearly independent, so there is only one zero column in $\mathring{\mathbf{M}}_{\mathcal{X}}$. Let the r-th column of $\mathring{\mathbf{M}}_{\mathcal{X}}$ be zero. Denote by $\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}}$ the matrix obtained by removing the r-th column from $\mathring{\mathbf{M}}_{\mathcal{X}}$, and similarly for $\mathbf{M}^{(-r)}_{\mathcal{X}}$. Let $\mathbf{Q}^1_{(-r)}$ be the vector obtained by removing the r-th entry from the vector $\mathbf{Q}^1$. According to Lemma 5, each n columns of $\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}}$ have rank at least $n-1$, so $\text{Rank}(\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}}) \leq n-1$. At the same time, $\text{Rank}_k(\mathbf{M}^{(-r)}_{\mathcal{X}}) = n$. Moreover, $\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}} \odot \mathbf{M}^{(-r)}_{\mathcal{X}}$ has $p-1$ columns. Hence $\text{Rank}(\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}}) + \text{Rank}_k(\mathbf{M}^{(-r)}_{\mathcal{X}}) \geq n-1+n = 2n-1 \geq (p-1)+1$ because it is assumed that $p \leq 2n-2$. Lemma 1 by Guo et al. [6] then implies that $\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}} \odot \mathbf{M}^{(-r)}_{\mathcal{X}}$ has full column rank. On the other hand, (23) becomes

  $$q_{r1} \cdot 0 + (\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}} \odot \mathbf{M}^{(-r)}_{\mathcal{X}})\mathbf{Q}^1_{(-r)} = (\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}} \odot \mathbf{M}^{(-r)}_{\mathcal{X}})\mathbf{Q}^1_{(-r)} = 0.$$

  Consequently, $\mathbf{Q}^1_{(-r)}$ is a zero vector because $\mathring{\mathbf{M}}^{(-r)}_{\mathcal{X}} \odot \mathbf{M}^{(-r)}_{\mathcal{X}}$ has full column rank. That is, only q_{r1} is non-zero. Then (22) tells us that

  $$\tilde{\mathbf{M}}^1_{\mathcal{X}} \otimes \tilde{\mathbf{M}}^1_{\mathcal{X}} = q_{r1} \cdot (\mathbf{M}^r_{\mathcal{X}} \otimes \mathbf{M}^r_{\mathcal{X}}).$$

  Hence, $\tilde{\mathbf{M}}^1_{\mathcal{X}}$ is a scaled version of $\mathbf{M}^r_{\mathcal{X}}$.

- Suppose no column of $\mathring{\mathbf{M}}_{\mathcal{X}}$ is zero. According to Lemma 5, each n columns of $\mathring{\mathbf{M}}_{\mathcal{X}}$ have rank at least $n-1$, so $\text{Rank}(\mathring{\mathbf{M}}_{\mathcal{X}}) \leq n-1$. Remember $\text{Rank}_k(\mathbf{M}^{(-r)}_{\mathcal{X}}) = n$ and $\mathring{\mathbf{M}}_{\mathcal{X}} \odot \mathbf{M}_{\mathcal{X}}$ has p columns. Hence $\text{Rank}(\mathring{\mathbf{M}}_{\mathcal{X}}) + \text{Rank}_k(\mathbf{M}_{\mathcal{X}}) \geq n-1+n = 2n-1 \geq p+1$ because $p \leq 2n-2$. Again, Lemma 1 by Guo et al. [6] indicates that $\mathring{\mathbf{M}}_{\mathcal{X}} \odot \mathbf{M}_{\mathcal{X}}$ has full column rank. Hence, in order for (23) to hold, $\mathbf{Q}^1$ must be a zero vector, leading to a contradiction.

Therefore, $\tilde{\mathbf{M}}^1_{\mathcal{X}}$ is a scaled version of $\mathbf{M}^r_{\mathcal{X}}$. Similarly $\tilde{\mathbf{M}}^2_{\mathcal{X}}$ is a scaled version of $\mathbf{M}^{r'}_{\mathcal{X}}$, and so on. Because $\tilde{\mathbf{M}}_{\mathcal{X}} \odot \tilde{\mathbf{M}}_{\mathcal{X}}$ has full column rank, different columns of $\tilde{\mathbf{M}}^1_{\mathcal{X}}$ must correspond to different columns of $\mathbf{M}_{\mathcal{X}}$. Further because of the symmetry between the two models (4) and (16), every column of $\tilde{\mathbf{M}}_{\mathcal{X}}$ must be proportional to a column of $\mathbf{M}_{\mathcal{X}}$ and vice versa.

□

1.2 Proof of Theorem 2

Theorem 2. *If* $\mathbf{F}$ *is identified up to trivialities, then the graph induced by* $\mathbf{F}$ *satisfies the bottleneck and strong non-redundancy conditions.*

We prove this for the two conditions separately. Throughout, we use $\mathbf{F}$, $\mathbf{M}$, and $\mathrm{Ch}(\cdot)$ to refer to the relevant components of the true causal system; and $\hat{\mathbf{F}}$, $\hat{\mathbf{M}}$, and $\mathrm{Ch}'(\cdot)$ to refer to the relevant components of an alternative causal system which we will construct. $\mathbf{R}$ is as described in Proposition 3.

Bottleneck condition: Let $B \neq \mathrm{Ch}(V_i)$ be a minimal bottleneck from $\mathrm{Ch}(V_i)$ to $\mathcal{X}$. Define

$$\hat{\mathbf{F}}^i = \mathbf{I}_{\mathcal{V}}^B \mathbf{R}_B^i$$

so that $\hat{\mathbf{F}}_j^i = \left[\mathbf{R}_B^i\right]_j$ if $V_j \in B$ and $\hat{\mathbf{F}}_j^i = 0$ otherwise. Clearly, $(\mathbf{I} - \hat{\mathbf{F}})$ is invertible whenever $(\mathbf{I} - \mathbf{F})$ is. Moreover,

$$\mathbf{M}_{\mathcal{X}} \hat{\mathbf{F}}^i = \mathbf{M}_{\mathcal{X}}^B \mathbf{R}_B^{\mathrm{Ch}(V_i)} \mathbf{F}_{\mathrm{Ch}(V_i)}^i = \mathbf{M}_{\mathcal{X}}^{\mathrm{Ch}(V_i)} \mathbf{F}_{\mathrm{Ch}(V_i)}^i = \mathbf{M}_{\mathcal{X}} \mathbf{F}^i.$$

So, since $\mathbf{M}_{\mathcal{X}} \hat{\mathbf{F}} = \mathbf{M}_{\mathcal{X}} \mathbf{F} = (\mathbf{M} - \mathbf{I})_{\mathcal{X}}$, (3) shows that $(\mathbf{I} - \hat{\mathbf{F}})_{\mathcal{X}}^{-1} = \mathbf{M}_{\mathcal{X}}$. Thus $\hat{\mathbf{F}}$ generates $\mathcal{M}$. Furthermore, $\|\hat{\mathbf{F}}\|_0 \leq \|\mathbf{F}\|_0$ since B is assumed to be minimal, so that $\hat{\mathbf{F}} \in \mathcal{F}$. Therefore, since $\mathbf{F}$ and $\hat{\mathbf{F}}$ induce different DAGs when $B \neq \mathrm{Ch}(V_i)$, $\mathbf{F}$ is not identified up to trivialities.

Parental non-redundancy: If $L_i \to V_j$, define $\mathbf{P}$ as the identity matrix with the i-th and j-th columns switched; $\mathbf{D}$ as the diagonal matrix with $D_i^i = 1/F_j^i$ and ones on the rest of the diagonal; and further

$$\begin{aligned}
\hat{\mathbf{M}}_{\mathcal{X}} &= \mathbf{M}_{\mathcal{X}} \mathbf{D} \mathbf{P}, \\
\hat{\mathbf{F}}^i &= \mathbf{I}^j + \mathbf{I}^i - \left(F_j^i\right)^{-1} \mathbf{P} \mathbf{F}^i, \\
\hat{\mathbf{F}}^j &= \mathbf{P} \left[\left(F_j^i\right)^{-1} \mathbf{F}^i + (\mathbf{F} - \mathbf{I})^j\right], \\
\hat{\mathbf{F}}^k &= \mathbf{P} \mathbf{D}^{-1} \mathbf{F}^k \text{ for all } k \notin \{i, j\},
\end{aligned}$$

so that $\mathrm{Ch}'(L_i) = \mathrm{Ch}(L_i)$ and $\mathrm{Ch}'(V_j) \subseteq \mathrm{Ch}(V_j)$, but with weights $\hat{\mathbf{F}}^i \not\propto \mathbf{F}^i$ and $\hat{\mathbf{F}}^j \not\propto \mathbf{F}^j$. Moreover, for every other V_k, if $L_i \in \mathrm{Ch}(V_k)$, then $V_j \in \mathrm{Ch}'(V_k)$ and vice versa. Clearly $\|\hat{\mathbf{F}}\|_0 \leq \|\mathbf{F}\|_0$ if L_i is a parental redundancy of V_j. Moreover, the resulting graph is acyclic whenever the true graph is acyclic. We compute:

$$\begin{aligned}
\hat{\mathbf{M}}_{\mathcal{X}} \hat{\mathbf{F}}^i &= \hat{\mathbf{M}}_{\mathcal{X}} \left[\mathbf{I}^j + \mathbf{I}^i - \left(F_j^i\right)^{-1} \mathbf{P} \mathbf{F}^i\right] \\
&= \mathbf{M}_{\mathcal{X}} \left(F_j^i\right)^{-1} \left[\mathbf{I}^i + F_j^i \mathbf{I}^j - \mathbf{F}^i\right] \\
&= \left(F_j^i\right)^{-1} \left[\mathbf{M}_{\mathcal{X}}^i + F_j^i \mathbf{M}_{\mathcal{X}}^j - \mathbf{M}_{\mathcal{X}}^i + \mathbf{I}_{\mathcal{X}}^i\right] \\
&= \mathbf{M}_{\mathcal{X}}^j \\
&= \hat{\mathbf{M}}_{\mathcal{X}}^i,
\end{aligned}$$

where we have used the fact that $\mathbf{I}_{\mathcal{X}}^i = 0$ since $L_i \in \mathcal{L}$. We further calculate

$$\begin{aligned}
\hat{\mathbf{M}}_{\mathcal{X}} \hat{\mathbf{F}}^j &= \mathbf{M} \mathbf{D} \left[\left(F_j^i\right)^{-1} \mathbf{F}^i + (\mathbf{F} - \mathbf{I})^j\right] \\
&= \left(F_j^i\right)^{-1} \mathbf{M}^i - \mathbf{I}^j \\
&= \hat{\mathbf{M}}^j - \mathbf{I}^j,
\end{aligned}$$

and finally

$$\hat{\mathbf{M}}_{\mathcal{X}} \hat{\mathbf{F}}^k = \mathbf{M}_{\mathcal{X}} \mathbf{F}^k$$

for $k \notin \{i, j\}$. Hence $\hat{\mathbf{M}}_{\mathcal{X}} \hat{\mathbf{F}} = (\mathbf{M} - \mathbf{I})_{\mathcal{X}}$. Rearranging, we see that $(\mathbf{I} - \hat{\mathbf{F}})_{\mathcal{X}}^{-1} = \hat{\mathbf{M}}_{\mathcal{X}} \in \mathcal{M}$. Because we have already argued that $\|\hat{\mathbf{F}}\|_0 \leq \|\mathbf{F}\|_0$ if L_i is a parental redundancy of V_j, it follows that $\hat{\mathbf{F}} \in \mathcal{F}$ if L_i is a parental redundancy of V_j. Therefore, due to the changes in causal scale, $\mathbf{F}$ is not identifiable up to trivialities.

Co-parental non-redundancy: If $L_i \not\to V_j$ and $V_k \in \mathrm{Ch}(L_i) \cap \mathrm{Ch}(V_j)$, define

$$\hat{\mathbf{F}}^j = \mathbf{F}^j + (\mathbf{I} - \mathbf{F})^i f,$$

where $f := \frac{F^j_k}{F^i_k}$. Then $\mathrm{Ch}'(V_j) = \mathrm{Ch}(V_j) \cup \{L_i\} \cup \mathrm{Ch}(L_i) - \{V_k\}$. If L_i is a co-parental redundancy of V_j, then the resulting system is no denser than the original. Moreover, the resulting system is acyclic. Finally, we calculate:

$$\begin{aligned}\mathbf{M}_{\mathcal{X}}\hat{\mathbf{F}}^j &= (\mathbf{M} - \mathbf{I})^j_{\mathcal{X}} + f\mathbf{I}^i_{\mathcal{X}} \\ &= (\mathbf{M} - \mathbf{I})^j_{\mathcal{X}},\end{aligned}$$

since $L_i \in \mathcal{L}$. With $\hat{\mathbf{F}}^k = \mathbf{F}^k$ for all remaining $k \neq i$, (3) shows that $(\hat{\mathbf{F}} - \mathbf{I})^{-1}_{\mathcal{X}} = \mathbf{M}_{\mathcal{X}}$, so that $\hat{\mathbf{F}} \in \mathcal{F}$. But since $\mathbf{F}$ induces a different DAG, $\mathbf{F}$ is not identified up to trivialities.

Remark: Notice that parental non-redundancy is not necessary to identify the full causal DAG; if L_i is a parental redundancy of V_j, and neither L_i nor V_j has any parent, then both the true system and the alternative system constructed in the proof of Theorem 2 will have the same skeleton. However, the alternative system is emphatically not a mere re-indexing and re-scaling of latent variables. In particular, if $V_j \in \mathcal{X}$, then the net effect of X_j on $\mathbf{X}$ will not be identified.

1.3 Characterizing bottleneck faithfulness

In this section we show that the set of adjacency matrices, corresponding to a fixed graph, that are not bottleneck faithful is a proper algebraic subset of all adjacency matrices for that graph. That is, the property of being bottleneck faithful is a generic property, both in the sense that it holds on an dense open set and that the exception set is of Lebesgue measure zero. But first, we prove Proposition 1 from the main paper.

Proof. (Prop. 1) Decompose $\mathbf{M}^J_K = \mathbf{M}^B_K \mathbf{R}^J_B$ using Proposition 3, then $\mathrm{Rank}\left(\mathbf{M}^J_K\right) \leq |B|$. □

To formalize that bottleneck faithfulness is a generic property, let G denote a graph (a DAG) with p nodes and n edges. A $p \times p$ adjacency matrix $\mathbf{F}$ that induces G has n nonzero entries, and we let $\mathbb{F}_G \subseteq \mathbb{R}^n$ denote the set of adjacency matrices that induce G – with $\mathbb{F}_G$ regarded as a subset of $\mathbb{R}^n$. An algebraic subset of $\mathbb{F}_G$ is a set

$$A = \{\mathbf{F} \in \mathbb{F}_G \mid \mathrm{pol}(\mathbf{F}) = 0\}$$

where pol is a polynomial. If pol is not the zero polynomial, A is a proper algebraic subset, and it is well known that A is then nowhere dense and of Lebesgue measure zero. We will construct a non-zero polynomial that evaluates to 0 if and only if the adjacency matrix is not bottleneck faithful. To this end, we first show that for any graph there exists an adjacency matrix that is bottleneck faithful.

Proposition 4. *For any graph G there exists $\mathbf{F} \in \mathbb{F}_G$ such that $\mathbf{F}$ is bottleneck faithful.*

Proof. The proof is by induction on the number of edges, n. Clearly for $n = 0$ we have bottleneck faithfulness.

For the induction step, let $n \geq 1$ and suppose there exists a bottleneck faithful adjacency matrix for any graph with less than n edges. Let G' be a graph with n edges, let V_i be a root node with $\mathrm{Ch}'(V_i) \neq \varnothing$ the children of V_i in G'. Let G be the subgraph of G' with all edges out of V_i removed, and let $\mathbf{F} \in \mathbb{F}_G$ denote a bottleneck faithful adjacency matrix for G. By choosing F^i_l for $l \in \mathrm{Ch}'(V_i)$ we can regard $\mathbf{F} \in \mathbb{F}_{G'}$, and the objective is to choose F^i_l such that $\mathbf{F}$ becomes bottleneck faithful for G'. In what follows, $\mathbf{M}$ denotes the mixing matrix for $\mathbf{F}$ on G, that is, when $F^i_l = 0$ for $l \in \mathrm{Ch}'(V_i)$, and $\mathbf{M}'$ denotes the mixing matrix on G' for any choice of F^i_l for $l \in \mathrm{Ch}'(V_i)$.

Given $J, K \subseteq \mathcal{V}$ we denote by L^J_K the set of coefficients F^i_l for $l \in \mathrm{Ch}'(V_i)$ for which bottleneck faithfulness for G' is **violated** by J and K.

If $V_i \notin J$, then since V_i is a root in G', bottleneck faithfulness holds for G' from J to K – no matter the coefficients F^i_l for $l \in \mathrm{Ch}'(V_i)$. Thus $L^J_K = \varnothing$.

If $V_i \in J$, we have that $\mathbf{M}^i_K = \mathbf{I}^i_K$, and

$$\mathbf{M}'^i_K = \mathbf{I}^i_K + \sum_{l \in \mathrm{Ch}'(V_i)} F^i_l \mathbf{M}^l_K \quad \text{and} \quad \mathbf{M}'^j_K = \mathbf{M}^j_K \text{ for } j \neq i.$$

There are two cases to consider.

1. *There is a minimal bottleneck, B, from J to K in G such that all paths from i to K in G' pass through B.*

   In this case, B is also a minimal bottleneck in G'. Since V_i is a root node, $\mathbf{M}_i^j = 0$ for all j, and replacing zero entries in the column $\mathbf{I}_K^i$ by possibly non-zero entries in the column $\mathbf{M}_K'^i$ cannot reduce the rank of $\mathbf{M}_K^J$. Thus no choice of F_l^i for $l \in \mathrm{Ch}'(V_i)$ makes J and K violate bottleneck faithfulness for G' and $L_K^J = \varnothing$.

2. *For any minimal bottleneck, B, from J to K in G there is a path from i to K in G' that does not pass through B.*

   Note that in this case, $V_i \notin K$ and $\mathbf{I}_K^i = 0$. Choose any minimal bottleneck, B, in G. Then $B \cup \{i\}$ is a minimal bottleneck in G', and with $\mathrm{col}(\mathbf{M}_K^J)$ the column space of $\mathbf{M}_K^J$,

$$L_K^J = \left\{ (F_l^i)_{l \in \mathrm{Ch}'(V_i)} \;\middle|\; \sum_{l \in \mathrm{Ch}'(V_i)} F_l^i \mathbf{M}_K^l \in \mathrm{col}(\mathbf{M}_K^J) \right\} \subseteq \mathbb{R}^{\mathrm{Ch}'(V_i)}.$$

   Note that L_K^J is a linear subspace of $\mathbb{R}^{\mathrm{Ch}'(V_i)}$. Due to bottleneck faithfulness for G, $\mathrm{Rank}(M_K^{J \cup \mathrm{Ch}'(V_i)}) \geq |B| + 1$, or there would be a bottleneck from J to K in G' of size $|B|$. This shows that L_K^J is a **true** subspace.

The set $\bigcap_{J,K} (L_K^J)^c$ contains all valid choices of coefficients F_l^i for $l \in \mathrm{Ch}'(V_i)$ such that $\mathbf{F}$ is bottleneck faithful for G'. Since all L_K^J are true subspaces, their complements are open and dense and so is their intersection. It is, in particular, non-empty and contains an element with $F_l^i \neq 0$ for all $l \in \mathrm{Ch}'(V_i)$. □

Proposition 5. *Let G be a graph with n edges. The set*

$$A = \{\mathbf{F} \in \mathbb{F}_G \mid \mathbf{F} \textit{ is bottleneck faithful for } G\}$$

is a proper algebraic subset of $\mathbb{R}^n$. In particular, a generic adjacency matrix $\mathbf{F}$ is bottleneck faithful.

Proof. Recall that the mixing matrix for $\mathbf{F}$ is

$$\mathbf{M} = (I - \mathbf{F})^{-1} = I + \mathbf{F} + \ldots + \mathbf{F}^p,$$

thus the entries in $\mathbf{M}$ are polynomials in the coefficients in $\mathbf{F}$. For any J and K, let b_K^J denote the size of a minimal bottleneck from J to K, let

$$\mathbb{H}_K^J = \{\mathbf{H} \mid \mathbf{H} \text{ is a } b_K^J \times b_K^J \text{ submatrix of } \mathbf{M}_K^J\},$$

and define

$$\mathrm{pol}_K^J(\mathbf{F}) = \sum_{\mathbf{H} \in \mathbb{H}_K^J} \det(\mathbf{H})^2.$$

Clearly, pol_K^J is a polynomial in the coefficients of $\mathbf{F}$, $\mathrm{pol}_K^J(\mathbf{F}) = 0$ if and only if bottleneck faithfulness is violated for $\mathbf{F}$ by J and K, and

$$A = \left\{ \mathbf{F} \in \mathbb{F}_G \;\middle|\; \prod_{J \subseteq \mathcal{V}, K \subseteq \mathcal{V}} \mathrm{pol}_K^J(\mathbf{F}) = 0 \right\}$$

is the set of adjacency matrices that are not bottleneck faithful. By Proposition 4 it is non-empty, thus the defining polynomial is not the zero polynomial and A is a proper algebraic subset. □

1.4 Proofs of Theorem 3 and its associated lemmas

First we prove a useful result not in the main text.

Proposition 6. *Suppose a partially observed DAG satisfies the bottleneck condition and generalized non-redundancy. For every V_i, V_j, $\mathrm{Ch}(V_i)$ is a bottleneck from $\mathrm{Ch}(V_j)$ to $\mathcal{X}$ if and only if $V_i = V_j$.*

Proof. The backward direction is obvious. Conversely, define

$$S := \{V_k \in \mathrm{Desc}(V_j) - \mathrm{Ch}(V_i) : \mathrm{Ch}(V_i) \text{ is a bottleneck from } V_k \text{ to } \mathcal{X}\}.$$

Obviously $S \cap \mathcal{X} = \varnothing$, and if $S = \varnothing$, then $\mathrm{Ch}(V_i)$ is not a bottleneck from V_j to X. Hence, by acyclicity, there is an $L_n \in S$ which has no descendent in S. Therefore $\mathrm{Ch}(L_n) \subseteq \mathrm{Ch}(V_i) \cup \{V_i\}$. □

As indicated in the main paper, we assume in Lemmas 1-4 and Theorem 3 that the bottleneck condition, strong non-redundancy, and bottleneck faithfulness hold, and that $\mathcal{M}$ is identifiable.

1.4.1 Proof of Lemma 1

Lemma 1. *Let $J = \mathrm{Ch}(V_i)$. Then the unique solution to (8) is given by $\mathbf{x} = \mathbf{F}^i_{\mathrm{Ch}(V_i)}$.*

Proof. For uniqueness, notice that

$$\mathrm{Rank}\left(\mathbf{M}_{\mathcal{X}}^{\mathrm{Ch}(V_i)}\right) \geq \mathrm{Rank}\left(\mathbf{M}_{\mathcal{X}-\{V_i\}}^{\mathrm{Ch}(V_i)}\right) = |\mathrm{Ch}(V_i)|,$$

with the equality following from the bottleneck condition and bottleneck faithfulness.

To see that $\mathbf{F}^i_J$ is a solution, notice that $\mathbf{M}^J_{\mathcal{X}}\mathbf{F}^i_J = \mathbf{M}_{\mathcal{X}}\mathbf{F}^i$, and use (3). □

1.4.2 Proof of Lemma 2

Lemma 2. *Suppose $J \subseteq \mathcal{V} - \mathrm{Anc}(V_i)$. If $(\mathbf{M} - \mathbf{I})^i_{\mathcal{X}} \in \mathrm{Range}\left(\mathbf{M}^J_{\mathcal{X}}\right)$, then $|J| \geq |\mathrm{Ch}(V_i)|$, with equality if and only if $J = \mathrm{Ch}(V_i)$.*

Proof. Let $J \subseteq \mathcal{V} - \mathrm{Anc}(V_i)$. The i-th row of (8) is satisfied trivially: $\mathbf{M}^J_i = 0$ when $J \subseteq \mathcal{V} - \mathrm{Anc}(V_i)$, and $(\mathbf{M} - \mathbf{I})^i_i = 0$ by acyclicity. Thus (8) has a solution if and only if

$$\mathbf{M}^i_{\mathcal{X}-\{V_i\}} = \mathbf{M}^J_{\mathcal{X}-\{V_i\}}\mathbf{x}$$

has a solution. This can be factorized as

$$\mathbf{M}^B_{\mathcal{X}-\{V_i\}}\mathbf{R}^i_B = \mathbf{M}^B_{\mathcal{X}-\{V_i\}}\mathbf{R}^J_B\mathbf{x},$$

where B is any minimal bottleneck from $\{V_i\} \cup J$ to $\mathcal{X} - \{V_i\}$. By bottleneck faithfulness,

$$\mathrm{Rank}\left(\mathbf{M}^B_{\mathcal{X}-\{V_i\}}\right) = |B|$$

so that there is a solution to (8) if and only if

$$\mathbf{R}^i_B = \mathbf{R}^J_B\mathbf{x}.$$

We distinguish two cases: either $V_i \in B$, or $V_i \notin B$.

In the first case, $\mathbf{R}^i_B$ is a B-dimensional basis vector with 1 in the i-th slot and 0 elsewhere. Thus if there is a solution, $\mathbf{R}^J_i \neq 0$, so that J has a path to V_i. Hence $J \cap \mathrm{Anc}(V_i) \neq \varnothing$.

In the second case, bottleneck faithfulness and the fact that B is a minimal bottleneck indicate that $|J| \geq |B|$. Noting that B is a bottleneck from $\mathrm{Ch}(V_i)$ to $\mathcal{X}$, we apply the bottleneck condition: $|B| \geq |\mathrm{Ch}(V_i)|$ with equality if and only if $B = \mathrm{Ch}(V_i)$. Moreover, for each $V_j \in J$, $\mathrm{Ch}(V_i)$ is a bottleneck from V_j to $\mathcal{X} - \{V_i\}$ if and only if $V_j \in \mathrm{Ch}(V_i)$ by Proposition 6. Combining, $|J| \geq |B| \geq |\mathrm{Ch}(V_i)|$ with equalities if and only if $J = B = \mathrm{Ch}(V_i)$. □

1.4.3 Proof of Lemma 3

Lemma 3. *For every $k \geq 0$, let $\mathcal{V}_k$ and $\mathcal{J}_k(V_i)$ be defined as in the main paper. Then $V_i \in \mathcal{V}_{k+1} - \mathcal{V}_k$ if and only if all of the following hold:*

1. *$V_i \notin \mathcal{V}_k$,*
2. *$|\mathrm{Support}(\mathbf{M}^i) - \mathcal{X}_k| \leq 1$,*

3. $|\mathcal{J}_{k+1}(V_i)| = 1$, *and*

4. *for all* $V_j \neq V_i$ *satisfying points 1 and 2,* $\mathbf{M}^j_{\mathcal{X}_k} \notin \text{Range}\left(\mathbf{M}^{J_k(V_i)}_{\mathcal{X}_k}\right)$.

Proof. Suppose $V_i \in \mathcal{V}_{k+1} - \mathcal{V}_k$. The first conjunct is obvious, the second conjunct follows by acyclicity, and the third conjunct follows from Lemma 2. For the fourth conjunct, take any other V_j, and let B be a minimal bottleneck from $\{V_j\} \cup \text{Ch}(V_i)$ to $\mathcal{X}_k$. Then B is further a bottleneck from $\{V_j\} \cup \text{Ch}(V_i)$ to $\mathcal{X}$. By Proposition 6, $B \neq \text{Ch}(V_i)$, since B is in particular a bottleneck from V_j to $\mathcal{X}$. Hence, by the bottleneck condition, $|B| > |\text{Ch}(V_i)|$, since B is in particular a bottleneck from $\text{Ch}(V_i)$ to X. Therefore,

$$\mathbf{R}^j_B = \mathbf{R}^{\text{Ch}(L_i)}_B \mathbf{x}$$

has no solution due to bottleneck faithfulness, so that

$$\mathbf{M}^j_{\mathcal{X}_k} = \mathbf{M}^{\text{Ch}(V_i)}_{\mathcal{X}_k} \mathbf{x}$$

also has no solution by bottleneck faithfulness applied to $\mathbf{M}^B_{\mathcal{X}_k}$.

Conversely, suppose $V_i \notin \mathcal{V}_{k+1} - \mathcal{V}_k$, and let $J \in \mathcal{J}_{k+1}(V_i)$. Clearly there exists some $V_j \in (\mathcal{V}_{k+1} - \mathcal{V}_k) \cap \text{Desc}(V_i)$. Now, for any minimal bottleneck B from $\{V_i\} \cup J$ to $\mathcal{X}_k$, $\mathbf{R}^i_B \in \text{Range}(\mathbf{R}^J_B)$ by bottleneck faithfulness on $\mathbf{M}^B_{\mathcal{X}_k}$. Since $J \subseteq \mathcal{V}_k$, it follows that $B \subseteq \mathcal{V}_k$; otherwise some row of $\mathbf{R}^J_B$ is zero, proving that either B is not minimal, or that J does not admit a solution to (8). So in particular, $\text{Anc}(V_j) \cap B = \varnothing$. Therefore, since V_i has a path to V_j, and since B is a bottleneck from V_i to $\mathcal{X}_k$, B is a bottleneck from V_j to $\mathcal{X}_k$. Hence

$$\mathbf{M}^j_{\mathcal{X}_k} \in \text{Range}\left(\mathbf{M}^B_{\mathcal{X}_k}\right) \subseteq \text{Range}\left(\mathbf{M}^J_{\mathcal{X}_k}\right),$$

since B is by definition a bottleneck from J to $\mathcal{X}_k$. Because V_j clearly satisfies conjuncts 1 and 2, the fourth conjunct is violated. □

1.4.4 Proof of Lemma 4

Lemma 4. $X_i \in \mathcal{X}_{k+1}$ *if and only if* $X_i \in \mathcal{V}_{k+1}$ *and* $\text{Support}\,(\mathbf{M}^i) - \mathcal{X}_k = \{i\}$.

Proof. This follows from definitions and acyclicity. □

1.4.5 Proof of Theorem 3

Theorem 3. *Suppose* $\mathbf{F}$ *satisfies strong non-redundancy, bottleneck faithfulness, and the bottleneck condition. Then* $\mathbf{F}$ *is identifiable up to trivialities.*

Proof. Lemma 3 shows that $\mathbf{M}^{\mathcal{V}_k}_{\mathcal{X}}$ is identifiable from $\mathcal{M}$ up to permutation and scaling of columns, since neither it nor Lemma 2 upon which it relies makes any assumptions about the scaling or permutation of $\mathbf{M}$.

From here, Lemma 4 shows that $\mathbf{M}^X_{\mathcal{X}}$ is identifiable up to scaling for each $X \in \mathcal{X}$. But then $\mathbf{M}^X_{\mathcal{X}}$ is identifiable exactly since $\mathbf{M}^X_X = 1$ by acyclicity. Hence $\mathbf{M}^{\mathcal{X}}_{\mathcal{X}}$ is identifiable exactly, and $\mathbf{M}^{\mathcal{L}}_{\mathcal{X}}$ up to permutation and scaling of columns. In other words,

$$\tilde{\mathcal{M}} := \{\mathbf{M}_{\mathcal{X}}\mathbf{D}\mathbf{P} : \mathbf{D}\mathbf{P} \in \mathcal{DP}_p \text{ with } (\mathbf{D}\mathbf{P})^{\mathcal{X}}_{\mathcal{X}} = \mathbf{I}\} \subset \mathcal{M}$$

is identifiable.

Fix any $\tilde{\mathbf{M}} \in \tilde{\mathcal{M}}$, and let $\mathbf{P}\mathbf{D}$ satisfy $\tilde{\mathbf{M}}\mathbf{P}\mathbf{D} = \mathbf{M}_{\mathcal{X}}$. Without loss of generality, reindex the latent variables so that $\mathbf{P} = \mathbf{I}$. Because $\mathcal{V}_k$ is identified for every k, apply Lemma 2 to conclude that $\tilde{\mathbf{M}}^{\text{Ch}(V_i)}$ is identifiable for every i. Moreover, notice that $\tilde{\mathbf{M}}^i = \tilde{\mathbf{M}}^{\text{Ch}(V_i)}\mathbf{x}$ if and only if $\mathbf{M}^i_{\mathcal{X}} = \mathbf{M}^{\text{Ch}(V_i)}_{\mathcal{X}}\left[\mathbf{D}^{\text{Ch}(V_i)}_{\text{Ch}(V_i)}\mathbf{x}/D^i_i\right]$. By Lemma 1, this holds if and only if $\mathbf{x} = D^i_i(\mathbf{D}^{-1})^{\text{Ch}(V_i)}_{\text{Ch}(V_i)}\mathbf{F}^i_{\text{Ch}(V_i)}$. Repeating for every V_i, $\tilde{\mathbf{F}} := \mathbf{D}^{-1}\mathbf{F}\mathbf{D}$ is identified from $\tilde{\mathbf{M}}$. □

2 Detailed experimental results

2.1 BIC penalty

As indicated in the main paper, we created 10 causal systems for each of the graphs in Figure 5. The causal weights were drawn uniformly from $(-0.9, -0.5) \cup (0.5, 0.9)$, and the independent variances $\sigma^2_{t,i}$ were drawn independently and uniformly from $(0.5, 2.0)$. For each of the T domains, n samples were then simulated according to (12).

To estimate the full system, we enumerate all partially observed DAGs with three observed and at most two latent variables, and optimize all causal weights and noise variances using L-BFGS. As initialization, every covariance term and non-zero causal weight was set to 1. We then selected the graph with the lowest BIC:

$$\mathrm{BIC}(\mathbf{F}, \Sigma) = -2\ell\ell(\mathbf{F}, \Sigma) + \|\mathbf{F}\|_0 \log(nT).$$

The table below summarizes the rate at which the correct skeleton was learned for each graph and each choice of $T \times n$.

Graph	1×5000	3×1666	5×500	5×1000	5×10000
(i)	10	10	10	10	10
(ii)	4	9	9	10	10
(iii)	2	10	9	8	9
(iv)	2	10	10	9	10
(v)	10	10	9	9	10
(vi)	1	7	4	6	8
(vii)	0	7	8	10	10
(viii)	0	7	8	9	10
(ix)	0	9	8	9	10
(x)	0	6	6	10	10
(xi)	0	9	6	9	10
(xii)	0	9	9	9	10
(xiii)	0	6	3	7	8
(xiv)	0	2	2	4	9
(xv)	0	0	1	3	6
(xvi)	2	9	8	9	10
(xvii)	0	0	1	3	5
(xviii)	0	5	7	8	10
(xix)	0	3	0	4	10
(xx)	0	5	2	6	10
(xxi)	0	3	1	2	8
(xxii)	0	5	0	5	8

Notice that the three sample sizes 1×5000, 3×1666, and 5×1000 all have the same total number of samples; any difference in performance is therefore attributable to the diversity of domains, and not to mere sample size.

Obviously these are not the only identifiable graphs; however, they are the only identifiable graphs up to re-indexing of variables. For example, graph (vii) has three versions: $X_1 \leftarrow L \rightarrow X_2$, $X_1 \leftarrow L \rightarrow X_3$, and $X_2 \leftarrow L \rightarrow X_3$. However, it is clearly sufficient to study the empirical recovery rate of only one of the three structures. Nevertheless, the exhaustive search was performed over all 1759 possible graphs with at most two latent variables; that is to say, for example, that the BIC optimization for graph (vii) included all three of these possibilities.

Recovery for graphs with exactly one latent and over 3 edges—graphs (xiii), (xiv), and (xv)—was relatively poor. In many incorrectly recovered graphs, the model of best fitting had an additional latent variable. In some sense, this gives the model an extra degree of freedom to approximate the covariances, by providing a larger overcomplete basis. However, only the number of edges was penalized in the L0 penalty, and not the number of latents. This was not possible in the case of (xii), as every graph with two latents has at least 4 edges, and so the BIC penalty was effective to prevent this. It is possible that this could be avoided by choosing the number of latent variables by a separate prior method, or by penalizing the number of latent variables. However, since this is not relevant

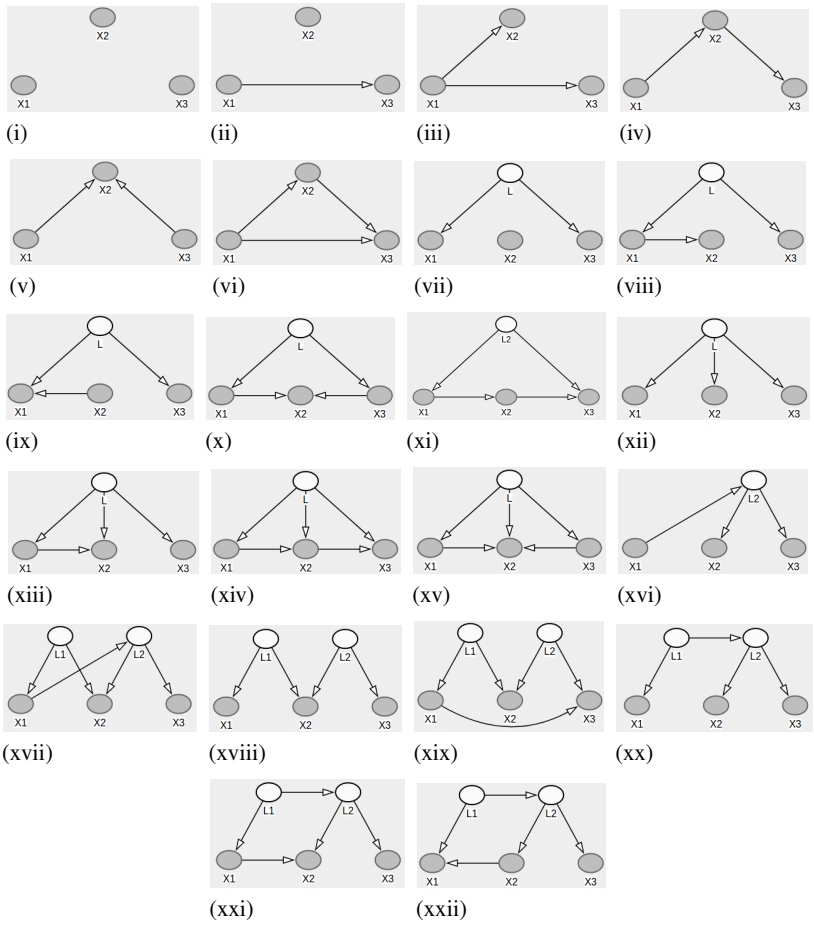


Figure 5: Identifiable graph skeletons, up to re-indexing of variables.

to enough graphs which are computationally admissible in an exhaustive search, we are not able to effectively study these cases.

The other graph with poor performance was (xvii). Note that in this case, with two latent variables and three measured variables, the mixing matrix $M_{\mathcal{X}}$ is not guaranteed to be recoverable by Theorem 1. (Recall that Theorem 1 gives sufficient identifiability conditions, which might not be necessary, and in the two cases the condition $p \leq 2n - 2$ does not hold.) However, in every alternative graph, the recovered system had a very different mixing matrix—even in terms of sparsity patterns up to permutation of columns! We therefore attribute this indeterminacy to a non-identifiable mixing matrix in the heterogeneous case, and not to an unidentifiable graph structure in general (for example, in the single-domain non-Gaussian setting).

For the sake of reproducibility, we have included the code for this main experiment, along with instructions for how to generate these results.

2.2 Detailed experimental results: Unidentifiable graphs

Here we show detailed estimation results for the three equivalence classes of Figure 6. For each of graphs (i) through (ix) of Figure 6, we generated 15 adjacency matrices as in the identifiable experiments—that is, with weights drawn uniformly from $(-0.9, -0.5) \cup (0.5, 0.9)$. Further, we drew 1000 samples from 5 domains, with the variance of each noise term $\sigma^2_{t,i}$ drawn uniformly from the interval $(0.5, 2.0)$. We then optimized the log likelihood for each of the three systems in each equivalence class, and selected the model with the best optimized log likelihood. Since we are only showing that the graphs in each class are equivalent, we do not need to search over all possible latent

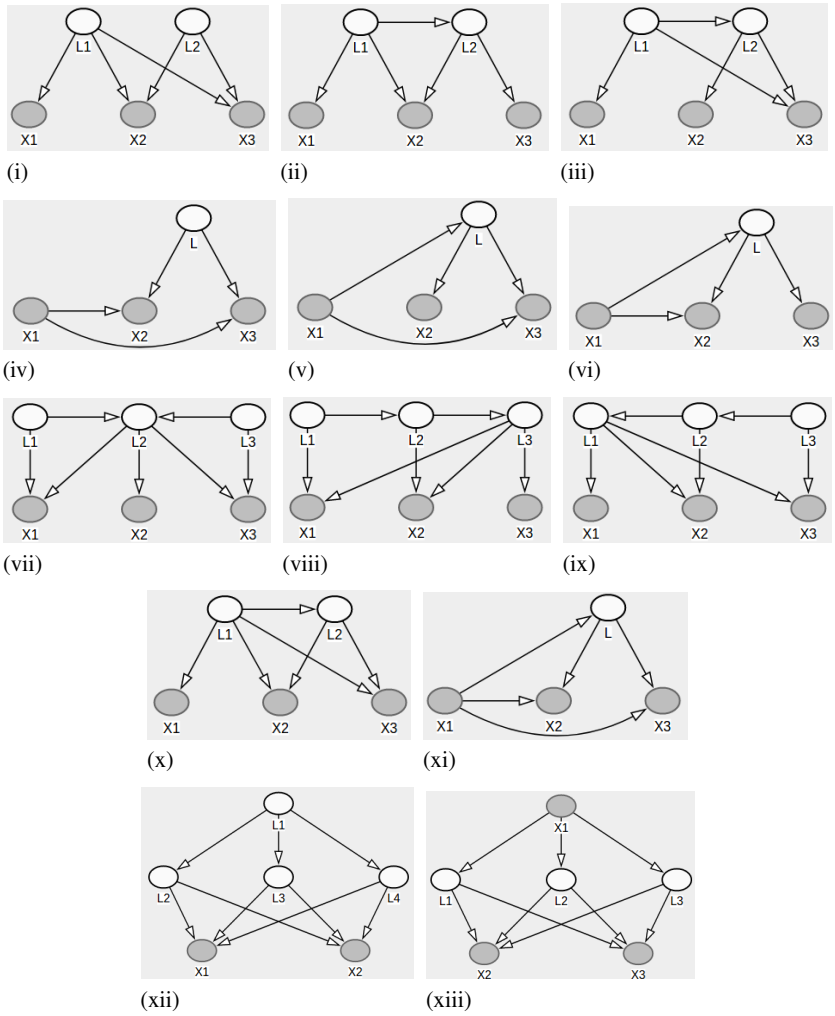


Figure 6: Three equivalence classes of graphs: {(i), (ii), (iii)} are equally sparse and observationally indistinguishable, as are {(iv), (v), (vi)} and {(vii), (viii), (ix)}. Moreover, (x) (xi) (xii) and (xiii) are not minimal, with (x) being observationally equivalent to (i) and (xi) being observationally equivalent to (iv). (xii) and (xiii) are discussed in Figure 1 of the main paper.

DAGs, but only over DAGs in the equivalence class. The charts below show the number of times each graph was chosen.

True graph	Times (i) was selected	Times (ii) was selected	Times (iii) was selected
(i)	6	3	6
(ii)	7	4	4
(iii)	7	1	7

True graph	Times (iv) was selected	Times (v) was selected	Times (vi) was selected
(iv)	8	3	4
(v)	6	4	5
(vi)	7	2	6

True graph	Times (vii) was selected	Times (viii) was selected	Times (ix) was selected
(vii)	2	5	8
(viii)	1	8	6
(ix)	1	7	7

In general, the difference in average log likelihood was on the same order as the convergence tolerance, $\|\nabla 2\ell\ell/n\|_\infty < 10^{-8}$, where n is the total number of samples.

We ran a similar experiment for the non-minimal graphs (x)-(xiii). For each, 10 heterogeneous causal systems were generated, and 1000 samples were simulated from each of the $T = 5$ domains. For every overly dense graph, on all 10/10 trials, a sparser and observationally equivalent graph received a lower BIC score than the true overly dense one.

2.3 Detailed experimental results: L1 penalty

In this section, we show exact results for a random system generated with skeleton (xviii). Results for other partially observed graphs are similar.

The true causal model is given by:

$$\begin{pmatrix} L_1 \\ L_2 \\ X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.82 & 0 & 0 & 0 & 0 \\ 0.53 & 0.51 & 0 & 0 & 0 \\ 0 & 0.82 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} L_1 \\ L_2 \\ X_1 \\ X_2 \\ X_3 \end{pmatrix} + \varepsilon.$$

As in the tests of the BIC algorithm, we used 5 domains. In the t-th domain, $\varepsilon \sim \mathcal{N}(0, \Sigma_t)$ for diagonal Σ_t. The variances for the t-th domain (i.e. the diagonal entries of Σ_t) are listed in the t-th row of the matrix below:

$$\mathbf{S} = \begin{bmatrix} 1.45 & 0.71 & 1.91 & 1.28 & 1.12 \\ 0.89 & 1.66 & 1.18 & 1.35 & 0.52 \\ 1.42 & 1.41 & 1.42 & 1.91 & 1.52 \\ 1.03 & 1.15 & 1.54 & 0.59 & 1.50 \\ 1.50 & 0.81 & 0.69 & 0.97 & 1.04 \end{bmatrix}.$$

All coefficients were randomly drawn by the same method used for the BIC-simulation studies.

We simulated 1000 observations for each of the 5 domains, and then estimated the adjacency matrix by minimizing

$$-2\ell\ell(\mathbf{F}, \Sigma)/(nT) + \lambda \sum |F_{i,j}|$$

as in (15), subject to $\sigma_{t,i} \in (0.1, 2.0)$ for each $t \in \{1, ..., 5\}$ and $i \in \mathcal{L}$. It is necessary to bound each of the latent σ, because otherwise it would be possible to evade the L1 penalty by making $\mathbf{F}^{\mathcal{L}}$ very small but $\Sigma^{\mathcal{L}}$ very large. However, to give the L1 optimizer the fairest chance of finding the true system, we constrained the latent values of σ with the same upper bound as Σ was generated with.

Because this is a non-convex objective, we ran L-BFGS-B from 10 random initializations, and used the point which best optimized (15).

Below we show the best-fitting adjacency matrix for various choices of λ. Recovered edges with strength in $(-0.1, 0.1)$ were pruned.

$\lambda = .5$; our procedure returned:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\lambda = .1$; our procedure returned:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.20 & 0 & 0 \\ 0 & 0 & 0 & 0.23 & 0 \end{bmatrix}$$

$\lambda = .015$; our procedure returned:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.37 & 0.24 & 0 & 0 \\ 0 & 0.31 & 0 & 0.19 & 0 \end{bmatrix}$$

Interestingly, at this choice of λ, the true graph was no longer a local minimum; with the truth as initialization, the optimizer moves to

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.45 & 0 & 0 & 0 & 0 \\ 0.39 & 0.36 & 0.12 & 0 & 0 \\ 0 & 0.41 & 0 & 0.16 & 0 \end{bmatrix}$$

which has a much smaller L1 penalty than the true system.

$\lambda = .01$; our procedure returned:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.39 & 0 & 0 & 0 & 0 \\ 0.38 & 0.37 & 0.16 & 0 & 0 \\ 0 & 0.32 & 0 & 0.21 & 0 \end{bmatrix}$$

With the truth as initialization, the optimizer moves to an adjacency matrix with similar support. Again, these systems incur a smaller L1 penalty than the true system, even though they are denser than the true system.

Similar results for each choice of λ are obtained with 10 000 samples in each domain.

References

[1] Kiyotoshi Matsuoka, Masahiro Ohoya, and Mitsuru Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[2] J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

[3] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

[4] J. B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Applicat.*, 18:95–138, 1977.

[5] R. Bro N. D. Sidiropoulos and G. B. Giannakis. Parallel factor analysis in sensor array processing. *IEEE Trans. Signal Processing*, 48:2377–2388, 2000.

[6] Xijing Guo, Sebastian Miron, David Brie, Shihua Zhu, and Xuewen Liao. A candecomp/parafac perspective on uniqueness of doa estimation using a vector sensor array. *IEEE Transactions on Signal Processing*, 59:3475–3481, 2011.

2.1 Additional Results and Discussion of Identifiability

In statistics, "identifiability" is usually defined in terms of the injectivity of maps from parameters to distributions. More formally, let $\mathcal{P}$ denote a set of probability distributions indexed by parameters θ and η so that $\mathcal{P} = \{P_{\theta,\eta}\}$. Throughout this section, we will say that a parameter θ_0 is **conventionally identifiable** if, for all θ, η, it is the case that $P_{\theta_0,\eta} = P_{\theta,\eta}$ if and only if $\theta_0 = \theta$. When identifiability holds for *all* θ_0, we say the parameter is globally identified; when it holds on the complement of a proper algebraic subset, it is globally identified.

This notion of conventional identifiability, widespread throughout statistics, is too strong for the purposes of [Identification], because the map from lower triangular adjacency matrices $\mathbf{F}$ to partially observed mixing matrices $\mathbf{M}_{\mathcal{X}}$ (and hence to the model class $\mathcal{P}$; see (4) of [Identification]) is never injective. For example, if an edge from V_i to V_j is increased by 1 and the vector of V_i's effects F^i is simultaneously decreased by F^j, then every variable will still have the same net effect on $\mathcal{X}$ so that $\mathcal{M}$ is unchanged.

If we want to talk about identifiability of the adjacency matrix for the partially observed linear model, one option is to introduce some form of razor, preference, or selection criterion to restrict the equivalence class from to a preferred subclass. We could then say a parameter of interest is "identifiable" when it is the only element of this preferred subclass. Formally, let $\theta \in \Theta$ denote the parameter of interest, and let $c : \Theta \to \mathbb{R}$ denote a cost function measuring the suitability of θ with respect to some preferred selection criterion. We say that a parameter θ_0 is c-**identifiable** if, for all θ and η, it is the case that $P_{\theta,\eta} = P_{\theta_0,\eta}$ with $c(\theta) \leq c(\theta_0)$ if and only if $\theta = \theta_0$. In Section 2.2, [Identification]defines "identifiability up to trivialities" in terms of minimality of $\mathbf{F}$; as such, [Identification] promises to investigate necessary and sufficient conditions for $\|\cdot\|_0$-identifiability of $\mathbf{F}$ (up to rescaling and re-indexing of latents). If we have a good reason to prefer sparser causal graphs, this could be a useful notion of identifiability to characterize.

While this is the notion of identifiability considered in [Identification] Theorem 2, it is not actually the notion of identifiability considered in [Identification] Theorem 3; there is a third way of caching out "identifiability" when conventional identifiability is impossible, and [Identification] Theorem 3 equivocates the two. Formally, let $Q : \mathcal{P} \to \mathbb{R}^m$ be a function from the model class of probability distributions to a finite dimensional vector space, and let $\mathtt{A}$ denote a polynomial time algorithm which requires $v \in \mathbb{R}^m$ and returns $\theta \in \Theta$. We say that θ_0 is $\mathtt{A}$-**identifiable** (from Q) if for all η it is the case that $\mathtt{A}(Q(P_{\theta_0,\eta})) = \theta_0$. The proof of [Identification] Theorem 3 uses Lemmas 1-4 to construct the true adjacency matrix $\mathbf{F}$ from the equivalence class of mixing matrices $\mathcal{M}$ whenever Conditions 1-3 are satisfied; call this implied polynomial time algorithm $\mathtt{A}_3$. As such, every adjacency matrix satisfying Conditions 1-3—and therefore every uniquely minimal graph—is $\mathtt{A}_3$-identifiable given $\mathcal{M}$. It is not obvious whether this is stronger or weaker than being $\|\cdot\|_0$-identifiable. Moreover, algorithmic identifiability is neither stronger nor weaker than conventional identifiability; a parameter may be provably conventionally identifiable but not constructively so, and the algorithm "always return 0" algorithmically identifies θ when $\theta = 0$ even if other values of θ are admissible.

To summarize: [Identification] Theorem 2 proves that every $\|\cdot\|_0$-identifiable adjacency

matrix satisfies Conditions 1 and 2, while [Identification] Theorem 3 proves that every adjacency matrix satisfying Conditions 1-3 is $\mathtt{A}_3$-identifiable; we have a merely necessary condition for one notion of identification, and a merely sufficient condition for the other.

To resolve this equivocation, we prove the following stronger result.

Theorem 1. *Suppose* $\mathbf{F}$ *satisfies Conditions 1-3. Then any other adjacency matrix* $\tilde{\mathbf{F}}$ *which generates* $\mathcal{M}_{\mathcal{X}}$ *violates bottleneck faithfulness.*

The proof uses [Identification] Lemma 2 and [Identification] Proposition 6 (printed in the supplement).

Proof. Suppose that $\tilde{\mathbf{F}}$ is bottleneck faithful, and define $\tilde{\mathcal{V}}_i$ and $\widetilde{\mathrm{Ch}}(\cdot)$ analogously to $\mathcal{V}_i$ and $\mathrm{Ch}(\cdot)$ from the main paper. We will show that $\mathbf{F} = \tilde{\mathbf{F}}$ by induction on the longest path length.

Trivially, $\tilde{\mathcal{V}}_1 \subseteq \mathcal{V}_1$. We indeed have $\tilde{\mathcal{V}}_1 = \mathcal{V}_1$ if $\tilde{\mathbf{F}}$ is bottleneck faithful.

Suppose for the purpose of induction that $\tilde{\mathcal{V}}_k = \mathcal{V}_k$ and $\widetilde{\mathrm{Ch}}(V_i) = \mathrm{Ch}(V_i)$ for every $V_i \in \mathcal{V}_k$. Now consider any $V_j \in \mathcal{V}_{k+1} - \mathcal{V}_k$. Because $\mathbf{F}$ satisfies Conditions 1 and 3, we know that $\operatorname{rank} \mathbf{M}_{\mathcal{X}-\{V_j\}}^{\mathrm{Ch}(V_j)\cup\{V_j\}} = |\mathrm{Ch}(V_j)|$, and so $\tilde{\mathbf{F}}$ must have a bottleneck from $\mathrm{Ch}(V_j) \cup \{V_j\}$ to $\mathcal{X} - \{V_j\}$ of size $|\mathrm{Ch}(V_j)|$ if it is to be bottleneck faithful. Notice that on $\mathbf{F}$, the unique smallest bottleneck from $\mathrm{Ch}(V_j)$ to $\mathcal{X} - \{V_j\}$ is $\mathrm{Ch}(V_j)$ itself due to Condition 1. But the same is true on $\tilde{\mathbf{F}}$; by definition $\widetilde{\mathrm{Ch}}(V_j) \subseteq \tilde{\mathcal{V}}_k = \mathcal{V}_k$, and $\mathbf{F}^{\mathcal{V}_k} = \tilde{\mathbf{F}}^{\mathcal{V}_k}$ by inductive hypothesis. Since any bottleneck of $\mathrm{Ch}(V_j) \cup \{V_j\}$ is also a bottleneck of $\mathrm{Ch}(V_j)$, it follows that $\mathrm{Ch}(V_j)$ is a bottleneck from V_j to $\mathcal{X} - \{V_j\}$ on $\tilde{\mathbf{F}}$.

Now, by analogy to the proof of [Identification] Proposition 6 (from the supplement), it can be shown that $\mathrm{Ch}(V_j)$ is not a bottleneck from any $V \notin \mathrm{Ch}(V_j) \cup \{V_j\}$ to $\mathcal{X} - \{V_j\}$ on $\mathbf{F}$. Hence $\mathbf{M}_{\mathcal{X}-\{V_j\}}^{j}$ is the only column of $\mathbf{M}_{\mathcal{X}-\{V_j\}}$ in the span of $\mathbf{M}_{\mathcal{X}-\{V_j\}}^{\mathrm{Ch}(V_j)}$ by bottleneck faithfulness on $\mathbf{F}$. This entails that no other $V \in \mathcal{V}$ can be bottlenecked from $\mathcal{X} - \{V_j\}$ by $\mathrm{Ch}(V_j)$ on $\tilde{\mathbf{F}}$, and thus that $\widetilde{\mathrm{Ch}}(V_j) \subseteq \mathrm{Ch}(V_j)$ (otherwise V_j would have a child not bottlenecked by $\mathrm{Ch}(V_j)$ on $\tilde{\mathbf{F}}$, contradicting the previous paragraph's conclusion). We in fact must have $\widetilde{\mathrm{Ch}}(V_j) = \mathrm{Ch}(V_j)$ by bottleneck faithfulness of $\mathbf{F}$. □

Theorem 1 shows that when bottleneck faithless adjacency matrices are excluded from the model a priori, Conditions 1-2 entail *conventional* identifiability *globally* on that submodel. Furthermore, since conventional identifiability obviously entails $\|\cdot\|_0$-identifiability, the converse also holds due to [Identification] Theorem 2. Hence, if the model class is restricted to bottleneck faithful adjacency matrices, then Conditions 1-2 are necessary and sufficient not only for $\|\cdot\|_0$-identifiability, but also for global conventional identifiability. We summarize this observation below.

Theorem 2. *Consider the model class described in Section 2 of [Identification] subject to the restriction that* $\mathbf{F}$ *satisfy bottleneck faithfulness. Then the following are equivalent:*

1. $\mathbf{F}$ *satisfies Conditions 1-2;*
2. $\mathbf{F}$ *is conventionally identifiable from* $\mathcal{M}$ *up to trivialities;*

3. $\mathbf{F}$ *is* $\|\cdot\|_0$*-identifiable.*

Let's be clear about what Theorems 1 and 2 are claiming. Although bottleneck faithfulness is generically satisfied on any *fixed* graph due to [Identification] Proposition 5 (from the supplement), Theorem 1 emphatically does *not* say that Conditions 1-2 entail generic (conventional) identifiable (up to trivialities); for every bottleneck faithful $\mathbf{F}$ with n edges, there exists an observationally equivalent $\mathbf{F}'$ with more than n edges which violates bottleneck faithfulness. What these theorems do claim is that if we are willing to exclude these bottleneck faithless alternatives from consideration a priori, then on this model class, Conditions 1-2 are equivalent to identifiability.

There are two potential justifications for the a priori exclusion of bottleneck faithless adjacency matrices. One is of course that bottleneck faithfulness violations constitute an proper algebraic subset of the parameter space for any graph over $\mathbf{V}$. Nature would only give us parameters from this particular null set if we were extremely unlucky. This is closely related to what Glymour et al. [1986] (page 94) call Spearman's Principle: "Other things being equal, prefer those models that, for all values of their free parameters, entail the constraints judged to hold in the populations." They elaborate: "Spearman's dominant methodological idea, never fully articulated, seems to have been that the best explanation [or model] is one that generates constraints found in the population measures without having to assume special values for its parameters." (page 236) Essentially, the graph over $\mathbf{V}$ entails certain rank constraints over submatrices of $\mathbf{M}_{\mathcal{X}}$—see [Identification] Proposition 1. So long as we have no reason to believe that the causal weights were fine-tuned to violate bottleneck faithfulness, excluding those models can be viewed as invoking Spearman's Principle with respect to the observed rank conditions on $\mathbf{M}_{\mathcal{X}}$.

Theorem 1 above is in one sense a very clear improvement over [Identification] Theorem 3; at the low cost of a priori excluding bottleneck faithless adjacency matrices from the model class, Theorem 1 gives us conventional identifiability—a much nicer notion than algorithmic identifiability, which (on its own) is merely relative to an arbitrary algorithm. But we should not forget about [Identification] Theorem 3. Conventional identifiability does not entail algorithmic identifiability; Theorem 1 is not constructive, and as such does not establish *algorithmic* identifiability. It is only through [Identification] Theorem 3 that we can conclude that the adjacency matrices treated in Theorem 2 are also algorithmically identifiable.

Similarly, we should not forget about [Identification] Theorem 2; since it holds globally with no reference to bottleneck faithfulness, it further motivates our study of the two graphical conditions.

When conventional identification is impossible, there are many weaker notions of identifiability available to be studied instead. Each can provide its own perspective on the identification problem at hand.

3 Multiple Causal Inference by Substitute Adjustment

This chapter contains the following paper:

[Adjustment] [Adams and Hansen, 2024]. J. Adams and N. R. Hansen. Substitute adjustment via recovery of latent variables. *arXiv preprint arXiv:2403.00202*, 2024.

In this chapter we study the multiple causal inference problem in the presence of unobserved confounding described in Section 1.1 (2).

Throughout the paper, many of the results and discussion are causally agnostic—the paper describes an adjustment problem without claiming that the regression model reflects anything about the interventional or counterfactual distributions. Nevertheless, the problem is intricately related to the multiple causal inference problem studied in Wang and Blei [2019]. The methods and theory presented there generated much discussion [D'Amour, 2019, Ogburn et al., 2020, Grimmer et al., 2023, for example]. In particular, much of the discussion was obscurred by questions about whether the deconfounder model of Wang and Blei [2019] is causally plausible. After the paper, we discuss the deconfounder project of Wang and Blei [2019] and how our results relate to the subsequent literature.

SUBSTITUTE ADJUSTMENT VIA RECOVERY OF LATENT VARIABLES

JEFFREY ADAMS[♯] AND NIELS RICHARD HANSEN[♯]

Abstract. The deconfounder was proposed as a method for estimating causal parameters in a context with multiple causes and unobserved confounding. It is based on recovery of a latent variable from the observed causes. We disentangle the causal interpretation from the statistical estimation problem and show that the deconfounder in general estimates adjusted regression target parameters. It does so by outcome regression adjusted for the recovered latent variable termed the substitute. We refer to the general algorithm, stripped of causal assumptions, as substitute adjustment. We give theoretical results to support that substitute adjustment estimates adjusted regression parameters when the regressors are conditionally independent given the latent variable. We also introduce a variant of our substitute adjustment algorithm that estimates an assumption-lean target parameter with minimal model assumptions. We then give finite sample bounds and asymptotic results supporting substitute adjustment estimation in the case where the latent variable takes values in a finite set. A simulation study illustrates finite sample properties of substitute adjustment. Our results support that when the latent variable model of the regressors hold, substitute adjustment is a viable method for adjusted regression.

1. Introduction

The deconfounder was proposed by Wang & Blei (2019) as a general algorithm for estimating causal parameters via outcome regression when: (1) there are multiple observed causes of the outcome; (2) the causal effects are potentially confounded by a latent variable; (3) the causes are conditionally independent given a latent variable Z. The proposal spurred discussion and criticism; see the comments to (Wang & Blei 2019) and the contributions by D'Amour (2019), Ogburn et al. (2020) and Grimmer et al. (2023). One question raised was whether the assumptions made by Wang & Blei (2019) are sufficient to claim that the deconfounder estimates a causal parameter. Though an amendment by Wang & Blei (2020) addressed the criticism and clarified their assumptions, it did not resolve all questions regarding the deconfounder.

The key idea of the deconfounder is to recover the latent variable Z from the observed causes and use this *substitute confounder* as a replacement for the unobserved confounder. The causal parameter is then estimated by outcome regression using the substitute confounder for adjustment. This way of adjusting for potential confounding has been in widespread use for some time in genetics and genomics, where, e.g., EIGENSTRAT based on PCA (Patterson et al. 2006, Price et al. 2006) was proposed to adjust for population structure in genome wide association studies (GWASs); see also (Song et al. 2015).

[♯]Department of Mathematical Sciences, University of Copenhagen
Universitetsparken 5, Copenhagen, 2100, Denmark
E-mail addresses: ja@math.ku.dk, niels.r.hansen@math.ku.dk.
Date: February 29, 2024.

Similarly, surrogate variable adjustment (Leek & Storey 2007) adjusts for unobserved factors causing unwanted variation in gene expression measurements.

In our view, the discussion regarding the deconfounder was muddled by several issues. First, issues with non-identifiablity of target parameters from the observational distribution with a *finite* number of observed causes lead to confusion. Second, the causal role of the latent variable Z and its causal relations to any unobserved confounder were difficult to grasp. Third, there was a lack of theory supporting that the deconfounder was actually estimating causal target parameters consistently. We defer the treatment of the thorny causal interpretation of the deconfounder to the discussion in Section 5 and focus here on the statistical aspects.

In our view, the statistical problem is best treated as *adjusted regression* without insisting on a causal interpretation. Suppose that we observe a real valued outcome variable Y and additional variables $X_1, X_2, \ldots, X_p$. We can then be interested in estimating the adjusted regression function

$$x \mapsto \mathbb{E}\left[\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}\right]\right] \tag{1}$$

where $\mathbf{X}_{-i}$ denotes all variables but X_i. That is, we adjust for all other variables when regressing Y on X_i. The adjusted regression function could have a causal interpretation in some contexts, but it is also of interest without a causal interpretation. It can, for instance, be used to study the added predictive value of X_i, and it is constant (as a function of x) if and only if $\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}\right] = \mathbb{E}\left[Y \mid \mathbf{X}_{-i}\right]$; that is, if and only if Y is conditionally mean independent of X_i given $\mathbf{X}_{-i}$ (Lundborg et al. 2023).

In the context of a GWAS, Y is a continuous phenotype and X_i represents a single nucleotide polymorphism (SNP) at the genomic site i. The regression function (1) quantifies how much a SNP at site i adds to the prediction of the phenotype outcome on top of all other SNP sites. In practice, only a fraction of all SNPs along the genome are observed, yet the number of SNPs can be in the millions, and estimation of the full regression model $\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i} = \mathbf{x}_{-i}\right]$ can be impossible without model assumptions. Thus if the regression function (1) is the target of interest, it is extremely useful if we, by adjusting for a substitute of a latent variable, can obtain a computationally efficient and statistically valid estimator of (1).

From our perspective, when viewing the problem as that of adjusted regression, the most pertinent questions are: (1) when is adjustment by the latent variable Z instead of $\mathbf{X}_{-i}$ appropriate; (2) can adjustment by substitutes of the latent variable, recovered from the observe X_i-s, be justified; (3) can we establish an asymptotic theory that allows for statistical inference when adjusting for substitutes?

With the aim of answering the three questions above, this paper makes two main contributions:

A transparent statistical framework. We focus on estimation of the adjusted mean, thereby disentangling the statistical problem from the causal discussion. This way the target of inference is clear and so are the assumptions we need about the observational distribution in terms of the latent variable model. We present in Section 2 a general framework with an infinite number of X_i-variables, and we present clear assumptions implying that we can replace adjustment by $\mathbf{X}_{-i}$ with adjustment by Z. Within the

general framework, we subsequently present an assumption-lean target parameter that is interpretable without restrictive model assumptions on the regression function.

A novel theoretical analysis. By restricting attention to the case where the latent variable Z takes values in a finite set, we give in Section 3 bounds on the estimation error due to using substitutes and on the recovery error—that is, the substitute mislabeling rate. These bounds quantify, among other things, how the errors depend on p; the actual (finite) number of X_i-s used for recovery. With minimal assumptions on the conditional distributions in the latent variable model and on the outcome model, we use our bounds to derive asymptotic conditions ensuring that the assumption-lean target parameter can be estimated just as well using substitutes as if the latent variables were observed.

To implement substitute adjustment in practice, we leverage recent developments on estimation in finite mixture models via tensor methods, which are computationally and statistically efficient in high dimensions. We illustrate our results via a simulation study in Section 4. Proofs and auxiliary results are in Appendix A. Appendix B contains a complete characterization of when recovery of Z is possible from an infinite $\mathbf{X}$ in a Gaussian mixture model.

1.1. Relation to existing literature.

Our framework and results are based on ideas by Wang & Blei (2019, 2020) and the literature preceding them on adjustment by surrogate/substitute variables. We add new results to this line of research on the theoretical justification of substitute adjustment as a method for estimation.

There is some literature on the theoretical properties of tests and estimators in high-dimensional problems with latent variables. Somewhat related to our framework is the work by Wang et al. (2017) on adjustment for latent confounders in multiple testing, motivated by applications to gene expression analysis. More directly related is the work by Ćevid et al. (2020) and Guo, Ćevid & Bühlmann (2022), who analyze estimators within a linear modelling framework with unobserved confounding. While their methods and results are definitely interesting, they differ from substitute adjustment, since they do not directly attempt to recover the latent variables. The linearity and sparsity assumptions, which we will not make, play an important role for their methods and analysis.

The paper by Grimmer et al. (2023) comes closest to our framework and analysis. Grimmer et al. (2023) present theoretical results and extensive numerical examples, primarily with a continuous latent variable. Their results are not favorable for the deconfounder and they conclude that the deconfounder is "not a viable substitute for careful research design in real-world applications". Their theoretical analyses are mostly in terms of computing the population (or n-asymptotic) bias of a method for a finite p (the number of X_i-variables), and then possibly investigate the limit of the bias as p tends to infinity. Compared to this, we analyze the asymptotic behaviour of the estimator based on substitute adjustment as n and p tend to infinity jointly. Moreover, since we specifically treat discrete latent variables, some of our results are also in a different framework.

2. Substitute adjustment

2.1. The General Model.

The full model is specified in terms of variables $(\mathbf{X}, Y)$, where $Y \in \mathbb{R}$ is a real valued outcome variable of interest and $\mathbf{X} \in \mathbb{R}^{\mathbb{N}}$ is a infinite vector of additional real valued variables. That is, $\mathbf{X} = (X_i)_{i \in \mathbb{N}}$ with $X_i \in \mathbb{R}$ for $i \in \mathbb{N}$. We let

$\mathbf{X}_{-i} = (X_j)_{j \in \mathbb{N} \setminus \{i\}}$, and define (informally) for each $i \in \mathbb{N}$ and $x \in \mathbb{R}$ the target parameter of interest

$$\chi_x^i = \mathbb{E}\left[\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}\right]\right]. \tag{2}$$

That is, χ_x^i is the mean outcome given $X_i = x$ when adjusting for all remaining variables $\mathbf{X}_{-i}$. Since $\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}\right]$ is generally not uniquely defined for all $x \in \mathbb{R}$ by the distribution of $(\mathbf{X}, Y)$, we need some additional structure to formally define χ_x^i. The following assumption and subsequent definition achieve this by assuming that a particular choice of the conditional expectation is made and remains fixed. Throughout, $\mathbb{R}$ is equipped with the Borel σ-algebra and $\mathbb{R}^{\mathbb{N}}$ with the corresponding product σ-algebra.

Assumption 1 (Regular Conditional Distribution)**.** Fix for each $i \in \mathbb{N}$ a Markov kernel $(P^i_{x,\mathbf{x}})_{(x,\mathbf{x}) \in \mathbb{R} \times \mathbb{R}^{\mathbb{N}}}$ on $\mathbb{R}$. Assume that $P^i_{x,\mathbf{x}}$ is the regular conditional distribution of Y given $(X_i, \mathbf{X}_{-i}) = (x, \mathbf{x})$ for all $x \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^{\mathbb{N}}$ and $i \in \mathbb{N}$. With P^{-i} the distribution of $\mathbf{X}_{-i}$, suppose additionally that

$$\iint |y| \, P^i_{x,\mathbf{x}}(\mathrm{d}y) P^{-i}(\mathrm{d}\mathbf{x}) < \infty$$

for all $x \in \mathbb{R}$.

Definition 1. Under Assumption 1 we define

$$\chi_x^i = \iint y \, P^i_{x,\mathbf{x}}(\mathrm{d}y) P^{-i}(\mathrm{d}\mathbf{x}). \tag{3}$$

Remark 1. Definition 1 makes the choice of conditional expectation explicit by letting

$$\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}\right] = \int y \, P^i_{x,\mathbf{X}_{-i}}(\mathrm{d}y)$$

be defined in terms of the specific regular conditional distribution that is fixed according to Assumption 1. We may need additional regularity assumptions to identify this Markov kernel from the distribution of $(\mathbf{X}, Y)$, which we will not pursue here.

The main assumption in this paper is the existence of a latent variable, Z, that will render the X_i-s conditionally independent, and which can be recovered from $\mathbf{X}$ in a suitable way. The variable Z will take values in a measurable space $(E, \mathcal{E})$, which we assume to be a Borel space. We use the notation $\sigma(Z)$ and $\sigma(\mathbf{X}_{-i})$ to denote the σ-algebras generated by Z and $\mathbf{X}_{-i}$, respectively.

Assumption 2 (Latent Variable Model)**.** There is a random variable Z with values in $(E, \mathcal{E})$ such that:

(1) $X_1, X_2, \ldots$ are conditionally independent given Z,
(2) $\sigma(Z) \subseteq \bigcap_{i=1}^{\infty} \sigma(\mathbf{X}_{-i})$.

The latent variable model given by Assumption 2 allows us to identify the adjusted mean by adjusting for the latent variable only.

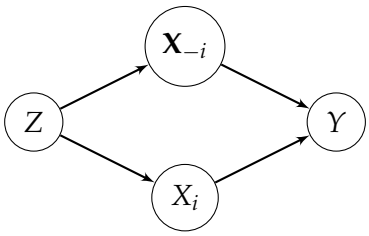


FIGURE 1. Directed Acyclic Graph (DAG) representing the joint distribution of $(X_i, \mathbf{X}_{-i}, Z, Y)$. The variable Z blocks the backdoor from X_i to Y.

Proposition 1. *Fix $i \in \mathbb{N}$ and let P_z^{-i} denote a regular conditional distribution of $\mathbf{X}_{-i}$ given $Z = z$. Under Assumptions 1 and 2, the Markov kernel*

$$Q^i_{x,z}(A) = \int P^i_{x,\mathbf{x}}(A) P_z^{-i}(\mathrm{d}\mathbf{x}), \qquad A \subseteq \mathbb{R} \tag{4}$$

is a regular conditional distribution of Y given $(X_i, Z) = (x, z)$, in which case

$$\chi^i_x = \iint y \, Q^i_{x,z}(\mathrm{d}y) P^Z(\mathrm{d}z) = \mathbb{E}\left[\mathbb{E}\left[Y \mid X_i = x; Z\right]\right]. \tag{5}$$

The joint distribution of $(X_i, \mathbf{X}_{-i}, Z, Y)$ is, by Assumption 2, Markov w.r.t. to the graph in Figure 1. Proposition 1 is essentially the backdoor criterion, since Z blocks the backdoor from X_i to Y via $\mathbf{X}_{-i}$; see Theorem 3.3.2 in (Pearl 2009) or Proposition 6.41(ii) in (Peters et al. 2017). Nevertheless, we include a proof in Appendix A for two reasons. First, Proposition 1 does not involve causal assumptions about the model, and we want to clarify that the mathematical result is agnostic to such assumptions. Second, the proof we give of Proposition 1 does not require regularity assumptions, such as densities of the conditional distributions, but it relies subtly on Assumption 2(2).

Example 1. Suppose $\mathbb{E}[|X_i|] \leqslant C$ for all i and some finite constant C, and assume, for simplicity, that $\mathbb{E}[X_i] = 0$. Let $\boldsymbol{\beta} = (\beta_i)_{i \in \mathbb{N}} \in \ell_1$ and define

$$\langle \boldsymbol{\beta}, \mathbf{X} \rangle = \sum_{i=1}^{\infty} \beta_i X_i.$$

The infinite sum converges almost surely since $\boldsymbol{\beta} \in \ell_1$. With ε being $\mathcal{N}(0,1)$-distributed and independent of $\mathbf{X}$ consider the outcome model

$$Y = \langle \boldsymbol{\beta}, \mathbf{X} \rangle + \varepsilon.$$

Letting $\boldsymbol{\beta}_{-i}$ denote the $\boldsymbol{\beta}$-sequence with the i-th coordinate removed, a straightforward, though slightly informal, computation, gives

$$\begin{aligned} \chi^i_x &= \mathbb{E}\left[\mathbb{E}\left[\beta_i X_i + \langle \boldsymbol{\beta}_{-i}, \mathbf{X}_{-i} \rangle \mid X_i = x; \mathbf{X}_{-i}\right]\right] \\ &= \beta_i x + \mathbb{E}\left[\langle \boldsymbol{\beta}_{-i}, \mathbf{X}_{-i} \rangle\right] = \beta_i x + \langle \boldsymbol{\beta}_{-i}, \mathbb{E}\left[\mathbf{X}_{-i}\right] \rangle = \beta_i x. \end{aligned}$$

To fully justify the computation, via Assumption 1, we let $P^i_{x,\mathbf{x}}$ be the $\mathcal{N}(\beta_i x + \langle \boldsymbol{\beta}_{-i}, \mathbf{x} \rangle, 1)$-distribution for the P^{-i}-almost all $\mathbf{x}$ where $\langle \boldsymbol{\beta}_{-i}, \mathbf{x} \rangle$ is well defined. For the remaining $\mathbf{x}$ we let $P^i_{x,\mathbf{x}}$ be the $\mathcal{N}(\beta_i x, 1)$-distribution. Then $P^i_{x,\mathbf{x}}$ is a regular conditional distribution

of Y given $(X_i, \mathbf{X}_{-i}) = (x, \mathbf{x})$,

$$\int y \, P^i_{x,\mathbf{x}}(\mathrm{d}y) = \beta_i x + \langle \boldsymbol{\beta}_{-i}, \mathbf{x} \rangle \quad \text{for } P^{-i}\text{-almost all } \mathbf{x},$$

and $\chi^i_x = \beta_i x$ follows from (3). It also follows from (4) that for P^Z-almost all $z \in E$,

$$\begin{aligned} \mathbb{E}\left[Y \mid X_i = x; Z = z\right] &= \int y \, Q^i_{x,z}(\mathrm{d}y) \\ &= \beta_i x + \int \langle \boldsymbol{\beta}_{-i}, \mathbf{x} \rangle P^{-i}_z(\mathrm{d}\mathbf{x}) \\ &= \beta_i x + \sum_{j \neq i} \beta_j \mathbb{E}[X_j \mid Z = z]. \end{aligned}$$

That is, with $\Gamma_{-i}(z) = \sum_{j \neq i} \beta_j \mathbb{E}[X_j \mid Z = z]$, the regression model

$$\mathbb{E}\left[Y \mid X_i = x; Z = z\right] = \beta_i x + \Gamma_{-i}(z)$$

is a partially linear model.

Example 2. While Example 1 is explicit about the outcome model, it does not describe an explicit latent variable model fulfilling Assumption 2. To this end, take $E = \mathbb{R}$, let $Z', U_1, U_2, \ldots$ be i.i.d. $\mathcal{N}(0,1)$-distributed and set $X_i = Z' + U_i$. By the Law of Large Numbers, for any $i \in \mathbb{N}$,

$$\frac{1}{n} \sum_{j=1; j \neq i}^{n+1} X_j = Z' + \frac{1}{n} \sum_{j=1; j \neq i}^{n+1} U_j \to Z'$$

almost surely for $n \to \infty$. Setting

$$Z = \begin{cases} \lim_{n \to \infty} \frac{1}{n} \sum_{j=1; j \neq i}^{n+1} X_j & \text{if the limit exists} \\ 0 & \text{otherwise} \end{cases}$$

we get that $\sigma(Z) \subseteq \sigma(\mathbf{X}_{-i})$ for any $i \in \mathbb{N}$ and $Z = Z'$ almost surely. Thus, Assumption 2 holds.

Continuing with the outcome model from Example 1, we see that for P^Z-almost all $z \in E$,

$$\mathbb{E}[X_j \mid Z = z] = \mathbb{E}[Z' + U_j \mid Z = z] = z,$$

thus $\Gamma_{-i}(z) = \gamma_{-i} z$ with $\gamma_{-i} = \sum_{j \neq i} \beta_j$. In this example it is actually possible to compute the regular conditional distribution, $Q^i_{x,z}$, of Y given $(X_i, Z) = (x, z)$ explicitly. It is the $\mathcal{N}\left(\beta_i x + \gamma_{-i} z, 1 + \|\boldsymbol{\beta}_{-i}\|_2^2\right)$-distribution where $\|\boldsymbol{\beta}_{-i}\|_2^2 = \langle \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}_{-i} \rangle$.

2.2. **Substitute Latent Variable Adjustment.** Proposition 1 tells us that under Assumptions 1 and 2 the adjusted mean, χ^i_x, defined by adjusting for the entire infinite vector $\mathbf{X}_{-i}$, is also given by adjusting for the latent variable Z. If the latent variable were observed we could estimate χ^i_x in terms of an estimate of the following regression function.

Definition 2 (Regression function)**.** Under Assumptions 1 and 2 define the regression function

$$b^i_x(z) = \int y \, Q^i_{x,z}(\mathrm{d}y) = \mathbb{E}\left[Y \mid X_i = x; Z = z\right] \tag{6}$$

where $Q^i_{x,z}$ is given by (4).

Algorithm 1: General Substitute Adjustment

1 **input:** data $\mathcal{S}_0 = \{\mathbf{x}^0_{1:p,1}, \ldots, \mathbf{x}^0_{1:p,m}\}$ and $\mathcal{S} = \{(\mathbf{x}_{1:p,1}, y_1), \ldots (\mathbf{x}_{1:p,n}, y_n)\}$, a set E, $i \in \{1, \ldots, p\}$ and $x \in \mathbb{R}$;

2 **options:** a method for estimating a recovery map $f^p : \mathbb{R}^p \to E$, a method for estimating the regression function $z \mapsto b^i_x(z)$;

3 **begin**

4 use data in $\mathcal{S}_0$ to compute the estimate $\hat{f}^p$ of the recovery map.

5 use data in $\mathcal{S}$ to compute the substitute latent variables as $\hat{z}_k := \hat{f}^p(\mathbf{x}_{1:p,k})$, $k = 1, \ldots, n$.

6 use data in $\mathcal{S}$ combined with the substitutes to compute the regression function estimate, $z \mapsto \hat{b}^i_x(z)$, and set

$$\widehat{\chi}^{i,\mathrm{sub}}_x = \frac{1}{n} \sum_{k=1}^{n} \hat{b}^i_x(\hat{z}_k).$$

7 **end**

8 **return** $\widehat{\chi}^{i,\mathrm{sub}}_x$

If we had n i.i.d. observations, $(x_{i,1}, z_1, y_1), \ldots, (x_{i,n}, z_n, y_n)$, of (X_i, Z, Y), a straightforward plug-in estimate of χ^i_x is

$$\hat{\chi}^i_x = \frac{1}{n} \sum_{k=1}^{n} \hat{b}^i_x(z_k), \tag{7}$$

where $\hat{b}^i_x(z)$ is an estimate of the regression function $b^i_x(z)$. In practice we do not observe the latent variable Z. Though Assumption 2(2) implies that Z can be recovered from $\mathbf{X}$, we do not assume we know this recovery map, nor do we in practice observe the entire $\mathbf{X}$, but only the first p coordinates, $\mathbf{X}_{1:p} = (X_1, \ldots, X_p)$.

We thus need an estimate of a recovery map, $\hat{f}^p : \mathbb{R}^p \to E$, such that for the *substitute latent variable* $\hat{Z} = \hat{f}^p(\mathbf{X}_{1:p})$ we have[1] that $\sigma(\hat{Z})$ approximately contains the same information as $\sigma(Z)$. Using such substitutes, a natural way to estimate χ^i_x is given by Algorithm 1, which is a general three-step procedure returning the estimate $\widehat{\chi}^{i,\mathrm{sub}}_x$.

The regression estimate $\hat{b}^i_x(z)$ in Algorithm 1 is computed on the basis of the substitutes, which likewise enter into the final computation of $\widehat{\chi}^{i,\mathrm{sub}}_x$. Thus the estimate is directly estimating $\chi^{i,\mathrm{sub}}_x = \mathbb{E}\left[\mathbb{E}\left[Y \mid X_i = x; \hat{Z}\right] \,\middle|\, \hat{f}^p\right]$, and it is expected to be biased as an estimate of χ^i_x. The general idea is that under some regularity assumptions, and for $p \to \infty$ and $m \to \infty$ appropriately, $\chi^{i,\mathrm{sub}}_x \to \chi^i_x$ and the bias vanishes asymptotically. Section 3 specifies a setup where such a result is shown rigorously.

Note that the estimated recovery map $\hat{f}^p$ in Algorithm 1 is the same for all $i = 1, \ldots, p$. Thus for any fixed i, the $x^0_{i,k}$-s are used for estimation of the recovery map, and the $x_{i,k}$-s are used for the computation of the substitutes. Steps 4 and 5 of the algorithm

[1] We can in general only hope to learn a recovery map of Z up to a Borel isomorphism, but this is also all that is needed, cf. Assumption 2.

could be changed to construct a recovery map $\hat{f}^p_{-i}$ independent of the i-th coordinate. This appears to align better with Assumption 2, and it would most likely make the $\hat{z}_k$-s slightly less correlated with the $x_{i,k}$-s. It would, on the other hand, lead to a slightly larger recovery error, and worse, a substantial increase in the computational complexity if we want to estimate $\widehat{\chi}^{i,\text{sub}}_x$ for all $i = 1, \ldots, p$.

Algorithm 1 leaves some options open. First, the estimation method used to compute $\hat{f}^p$ could be based on any method for estimating a recovery map, e.g., using a factor model if $E = \mathbb{R}$ or a mixture model if E is finite. The idea of such methods is to compute a parsimonious $\hat{f}^p$ such that: (1) conditionally on $\hat{z}^0_k = \hat{f}^p(\mathbf{x}^0_{1:p,k})$ the observations $x^0_{1,k}, \ldots, x^0_{p,k}$ are approximately independent for $k = 1, \ldots, m$; and (2) $\hat{z}^0_k$ is minimally predictive of $x^0_{i,k}$ for $i = 1, \ldots, p$. Second, the regression method for estimation of the regression function $b^i_x(z)$ could be any parametric or nonparametric method. If $E = \mathbb{R}$ we could use OLS combined with the parametric model $b^i_x(z) = \beta_0 + \beta_i x + \gamma_{-i} z$, which would lead to the estimate

$$\widehat{\chi}^{i,\text{sub}}_x = \hat{\beta}_0 + \hat{\beta}_i x + \hat{\gamma}_{-i} \frac{1}{n} \sum_{k=1}^{n} \hat{z}_k.$$

If E is finite, we could still use OLS but now combined with the parametric model $b^i_x(z) = \beta'_{i,z} x + \gamma_{-i,z}$, which would lead to the estimate

$$\widehat{\chi}^{i,\text{sub}}_x = \left(\frac{1}{n} \sum_{k=1}^{n} \hat{\beta}'_{i,\hat{z}_k} \right) x + \frac{1}{n} \sum_{k=1}^{n} \hat{\gamma}_{-i,\hat{z}_k}.$$

The relation between the two datasets in Algorithm 1 is not specified by the algorithm either. It is possible that they are independent, e.g., by data splitting, in which case $\hat{f}^p$ is independent of the data in $\mathcal{S}$. It is also possible that $m = n$ and $\mathbf{x}^0_{1:p,k} = \mathbf{x}_{1:p,k}$ for $k = 1, \ldots, n$. While we will assume $\mathcal{S}_0$ and $\mathcal{S}$ independent for the theoretical analysis, the $\mathbf{x}_{1:p}$-s from $\mathcal{S}$ will in practice often be part of $\mathcal{S}_0$, if not all of $\mathcal{S}_0$.

2.3. **Assumption-Lean Substitute Adjustment.** If the regression model in the general Algorithm 1 is misspecified we cannot expect that $\widehat{\chi}^{i,\text{sub}}_x$ is a consistent estimate of χ^i_x. In Section 3 we investigate the distribution of a substitute adjustment estimator in the case where E is finite. It is possible to carry out this investigation assuming a partially linear regression model, $b^i_x(z) = \beta_i x + \Gamma_{-i}(z)$, but the results would then hinge on this model being correct. To circumvent such a model assumption we proceed instead in the spirit of *assumption-lean regression* (Berk et al. 2021, Vansteelandt & Dukes 2022). Thus we focus on a univariate target parameter defined as a functional of the data distribution, and we then investigate its estimation via substitute adjustment.

Assumption 3 (Moments)**.** It holds that $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}[X_i^2] < \infty$ and $\mathbb{E}\left[\text{Var}\left[X_i \mid Z\right]\right] > 0$.

Definition 3 (Target parameter)**.** Let $i \in \mathbb{N}$. Under Assumptions 2 and 3 define the target parameter

$$\beta_i = \frac{\mathbb{E}\left[\text{Cov}\left[X_i, Y \mid Z\right]\right]}{\mathbb{E}\left[\text{Var}\left[X_i \mid Z\right]\right]}. \tag{8}$$

Algorithm 2 gives a procedure for estimating β_i based on substitute latent variables. The following proposition gives insight on the interpretation of the target parameter β_i.

Algorithm 2: Assumption-Lean Substitute Adjustment

input: data $\mathcal{S}_0 = \{\mathbf{x}^0_{1:p,1}, \ldots, \mathbf{x}^0_{1:p,m}\}$ and $\mathcal{S} = \{(\mathbf{x}_{1:p,1}, y_1), \ldots (\mathbf{x}_{1:p,n}, y_n)\}$, a set E and $i \in \{1, \ldots, p\}$;

options: a method for estimating the recovery map $f^p : \mathbb{R}^p \to E$, methods for estimating the regression functions $\mu_i(z) = \mathbb{E}[X_i \mid Z = z]$ and $g(z) = \mathbb{E}[Y \mid Z = z]$;

begin

use data in $\mathcal{S}_0$ to compute the estimate $\hat{f}^p$ of the recovery map.

use data in $\mathcal{S}$ to compute the substitute latent variables as $\hat{z}_k := \hat{f}^p(\mathbf{x}_{1:p,k})$, $k = 1, \ldots, n$.

use data in $\mathcal{S}$ combined with the substitutes to compute the regression function estimates $z \mapsto \hat{\mu}_i(z)$ and $z \mapsto \hat{g}(z)$, and set

$$\widehat{\beta}_i^{\text{sub}} = \frac{\sum_{k=1}^n (x_{i,k} - \hat{\mu}_i(\hat{z}_k))(y_k - \hat{g}(\hat{z}_k))}{\sum_{k=1}^n (x_{i,k} - \hat{\mu}_i(\hat{z}_k))^2}.$$

end

return $\widehat{\beta}_i^{\text{sub}}$

Proposition 2. *Under Assumptions 1, 2 and 3, and with $b^i_x(z)$ given as in Definition 2, and β_i given as in Definition 3,*

$$\beta_i = \frac{\mathbb{E}\left[\text{Cov}\left[X_i, b^i_{X_i}(Z) \mid Z\right]\right]}{\mathbb{E}\left[\text{Var}\left[X_i \mid Z\right]\right]}. \tag{9}$$

Moreover, $\beta_i = 0$ if $b^i_x(z)$ does not depend on x. If $b^i_x(z) = \beta_i'(z)x + \Gamma_{-i}(z)$ then

$$\beta_i = \mathbb{E}\left[w_i(Z)\beta_i'(Z)\right] \tag{10}$$

where

$$w_i(Z) = \frac{\text{Var}[X_i \mid Z]}{\mathbb{E}\left[\text{Var}[X_i \mid Z]\right]}.$$

We include a proof of Proposition 2 in Appendix A.1 for completeness. The arguments are essentially as in (Vansteelandt & Dukes 2022).

Remark 2. If $b^i_x(z) = \beta_i'(z)x + \Gamma_{-i}(z)$ it follows from Proposition 1 that $\chi^i_x = \beta_i' x$, where the coefficient $\beta_i' = \mathbb{E}[\beta_i'(Z)]$ may differ from β_i given by (10). In the special case where the variance of X_i given Z is constant across all values of Z, the weights in (10) are all 1, in which case $\beta_i = \beta_i'$. For the partially linear model, $b^i_x(z) = \beta_i' x + \Gamma_{-i}(z)$, with β_i' not depending on z, it follows from (10) that $\beta_i = \beta_i'$ irrespectively of the weights.

Remark 3. If $X_i \in \{0, 1\}$ then $b^i_x(z) = (b^i_1(Z) - b^i_0(Z))x + b^i_0(Z)$, and the contrast $\chi^i_1 - \chi^i_0 = \mathbb{E}\left[b^i_1(Z) - b^i_0(Z)\right]$ is an unweighted mean of differences, while it follows from (10) that

$$\beta_i = \mathbb{E}\left[w_i(Z)(b^i_1(Z) - b^i_0(Z))\right]. \tag{11}$$

If we let $\pi_i(Z) = \mathbb{P}(X_i = 1 \mid Z)$, we see that the weights are given as

$$w_i(Z) = \frac{\pi_i(Z)(1 - \pi_i(Z))}{\mathbb{E}\left[\pi_i(Z)(1 - \pi_i(Z))\right]}.$$

We summarize three important take-away messages from Proposition 2 and the remarks above as follows:

Conditional mean independence: The null hypothesis of conditional mean independence,

$$\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i}]\right) = \mathbb{E}\left[Y \mid \mathbf{X}_{-i}\right],$$

implies that $\beta_i = 0$. The target parameter β_i thus suggests an assumption-lean approach to testing this null without a specific model of the conditional mean.

Heterogeneous partial linear model: If the conditional mean,

$$b_x^i(z) = \mathbb{E}\left[Y \mid X_i = x; Z = z\right],$$

is linear in x with an x-coefficient that depends on Z (heterogeneity), the target parameter β_i is a *weighted* mean of these coefficients, while $\chi_x^i = \beta_i' x$ with β_i' the *unweighted* mean.

Simple partial linear model: If the conditional mean is linear in x with an x-coefficient that is *independent* of Z (homogeneity), the target parameter β_i coincides with this x-coefficient and $\chi_x^i = \beta_i x$. Example 1 is a special case where the latent variable model is arbitrary but the full outcome model is linear.

Just as for the general Algorithm 1, the estimate that Algorithm 2 outputs, $\widehat{\beta}_i^{\text{sub}}$, is not directly estimating the target parameter β_i. It is directly estimating

$$\beta_i^{\text{sub}} = \frac{\mathbb{E}\left[\text{Cov}\left[X_i, Y \mid \hat{Z}\right] \,\middle|\, \hat{f}^p\right]}{\mathbb{E}\left[\text{Var}\left[X_i \mid \hat{Z}\right] \,\middle|\, \hat{f}^p\right]}. \tag{12}$$

Fixing the estimated recovery map $\hat{f}^p$ and letting $n \to \infty$, we can expect that $\widehat{\beta}_i^{\text{sub}}$ is consistent for β_i^{sub} and not for β_i.

Pretending that the z_k-s were observed, we introduce the oracle estimator

$$\widehat{\beta}_i = \frac{\sum_{k=1}^n (x_{i,k} - \overline{\mu}_i(z_k))(y_k - \overline{g}(z_k))}{\sum_{k=1}^n (x_{i,k} - \overline{\mu}_i(z_k))^2}.$$

Here, $\overline{\mu}_i$ and $\overline{g}$ denote estimates of the regression functions μ_i and g, respectively, using the z_k-s instead of the substitutes. The estimator $\widehat{\beta}_i$ is independent of m, p, and $\hat{f}^p$, and when $(x_{i,1}, z_1, y_1), \ldots, (x_{i,n}, z_n, y_n)$ are i.i.d. observations, standard regularity assumptions (van der Vaart 1998) will ensure that the estimator $\widehat{\beta}_i$ is consistent for β_i (and possibly even $\sqrt{n}$-rate asymptotically normal). Writing

$$\widehat{\beta}_i^{\text{sub}} - \beta_i = (\widehat{\beta}_i^{\text{sub}} - \widehat{\beta}_i) + (\hat{\beta}_i - \beta_i) \tag{13}$$

we see that if we can appropriately bound the error, $|\widehat{\beta}_i^{\text{sub}} - \widehat{\beta}_i|$, due to using the substitutes instead of the unobserved z_k-s, we can transfer asymptotic properties of $\hat{\beta}_i$ to $\widehat{\beta}_i^{\text{sub}}$. It is the objective of the following section to demonstrate how such a bound can be achieved for a particular model class.

3. Substitute adjustment in a mixture model

In this section, we present a theoretical analysis of assumption-lean substitute adjustment in the case where the latent variable takes values in a finite set. We provide finite-sample bounds on the error of $\widehat{\beta}_i^{\text{sub}}$ due to the use of substitutes, and we show, in particular, that there exist trajectories of m, n and p along which the estimator is asymptotically equivalent to the oracle estimator $\widehat{\beta}_i$, which uses the actual latent variables.

3.1. The mixture model.

To be concrete, we assume that $\mathbf{X}$ is generated by a finite mixture model such that conditionally on a latent variable Z with values in a finite set, the coordinates of $\mathbf{X}$ are independent. The precise model specification is as follows.

Assumption 4 (Mixture Model)**.** There is a latent variable Z with values in the finite set $E = \{1, \ldots, K\}$ such that $X_1, X_2, \ldots$ are conditionally independent given $Z = z$. Furthermore,

(1) The conditional distribution of X_i given $Z = z$ has finite second moment, and its conditional mean and variance are denoted

$$\mu_i(z) = \mathbb{E}[X_i \mid Z = z]$$
$$\sigma_i^2(z) = \text{Var}[X_i \mid Z = z]$$

for $z \in E$ and $i \in \mathbb{N}$.

(2) The conditional means satisfy the following *separation* condition

$$\sum_{i=1}^{\infty} (\mu_i(z) - \mu_i(v))^2 = \infty \tag{14}$$

for all $z, v \in E$ with $v \neq z$.

(3) There are constants $0 < \sigma_{\min}^2 \leqslant \sigma_{\max}^2 < \infty$ that bound the conditional variances;

$$\sigma_{\min}^2 \leqslant \max_{z \in E} \sigma_i^2(z) \leqslant \sigma_{\max}^2 \tag{15}$$

for all $i \in \mathbb{N}$.

(4) $\mathbb{P}(Z = z) > 0$ for all $z \in E$.

Algorithm 3 is one specific version of Algorithm 2 for computing $\hat{\beta}_i^{\text{sub}}$ when the latent variable takes values in a finite set E. The recovery map in Step 5 is given by computing the nearest mean, and it is thus estimated in Step 4 by estimating the means for each of the mixture components. How this is done precisely is an option of the algorithm. Once the substitutes are computed, outcome means and $x_{i,k}$-means are (re)computed within each component. The computations in Steps 6 and 7 of Algorithm 3 result in the same estimator as the OLS estimator of β_i when it is computed using the linear model

$$b_x^i(z) = \beta_i x + \gamma_{-i,z}, \qquad \beta_i, \gamma_{-i,1}, \ldots, \gamma_{-i,K} \in \mathbb{R}$$

on the data $(x_{i,1}, \hat{z}_1, y_1), \ldots (x_{i,n}, \hat{z}_n, y_n)$. This may be relevant in practice, but it is also used in the proof of Theorem 1. The corresponding oracle estimator, $\hat{\beta}_i$, is similarly an OLS estimator.

Algorithm 3: Assumption Lean Substitute Adjustment w. Mixtures

input: data $\mathcal{S}_0 = \{\mathbf{x}^0_{1:p,1}, \ldots, \mathbf{x}^0_{1:p,m}\}$ and $\mathcal{S} = \{(\mathbf{x}_{1:p,1}, y_1), \ldots (\mathbf{x}_{1:p,n}, y_n)\}$, a finite set E and $i \in \{1, \ldots, p\}$;

options: a method for estimating the conditional means $\mu_j(z) = \mathbb{E}[X_j \mid Z = z]$;

begin

use the data in $\mathcal{S}_0$ to compute the estimates $\check{\mu}_j(z)$ for $j \in \{1, \ldots, p\}$ and $z \in E$.

use the data in $\mathcal{S}$ to compute the substitute latent variables as $\hat{z}_k = \arg\min_z \|\mathbf{x}_{1:p,k} - \check{\boldsymbol{\mu}}_{1:p}(z)\|_2$, $k = 1, \ldots, n$.

use the data in $\mathcal{S}$ combined with the substitutes to compute the estimates

$$\hat{g}(z) = \frac{1}{\hat{n}(z)} \sum_{k:\hat{z}_k = z} y_k, \quad z \in E$$

$$\hat{\mu}_i(z) = \frac{1}{\hat{n}(z)} \sum_{k:\hat{z}_k = z} x_{i,k}, \quad z \in E,$$

where $\hat{n}(z) = \sum_{k=1}^n 1(\hat{z}_k = z)$ is the number of k-s with $\hat{z}_k = z$.

use the data in $\mathcal{S}$ combined with the substitutes to compute

$$\widehat{\beta}_i^{\text{sub}} = \frac{\sum_{k=1}^n (x_{i,k} - \hat{\mu}_i(\hat{z}_k))(y_k - \hat{g}(\hat{z}_k))}{\sum_{k=1}^n (x_{i,k} - \hat{\mu}_i(\hat{z}_k))^2}.$$

end

return $\widehat{\beta}_i^{\text{sub}}$

Note that Assumption 4 implies that

$$\mathbb{E}[X_i^2] = \sum_{z \in E} \mathbb{E}[X_i^2 \mid Z = z]\mathbb{P}(Z = z) = \sum_{z \in E} (\sigma_i^2(z) + \mu_i(z)^2)\mathbb{P}(Z = z) < \infty$$

$$\mathbb{E}\left[\text{Var}\left[X_i \mid Z\right]\right] = \sum_{z \in E} \sigma_i^2(z)\mathbb{P}(Z = z) \geqslant \sigma^2_{\min} \min_{z \in E} \mathbb{P}(Z = z) > 0.$$

Hence Assumption 4, combined with $\mathbb{E}[Y^2] < \infty$, ensure that the moment conditions in Assumption 3 hold.

The following proposition states that the mixture model given by Assumption 4 is a special case of the general latent variable model.

Proposition 3. *Assumption 4 on the mixture model implies Assumption 2. Specifically, that* $\sigma(Z) \subseteq \sigma(\mathbf{X}_{-i})$ *for all* $i \in \mathbb{N}$.

Remark 4. The proof of Proposition 3 is in Appendix A.3. Technically, the proof only gives *almost sure* recovery of Z from $\mathbf{X}_{-i}$, and we can thus only conclude that $\sigma(Z)$ is contained in $\sigma(\mathbf{X}_{-i})$ up to negligible sets. We can, however, replace Z by a variable, Z', such that $\sigma(Z') \subseteq \sigma(\mathbf{X}_{-i})$ and $Z' = Z$ almost surely. We can thus simply swap Z with Z' in Assumption 4.

Remark 5. The arguments leading to Proposition 3 rely on Assumptions 4(2) and 4(3)—specifically the separation condition (14) and the upper bound in (15). However, these conditions are not necessary to be able to recover Z from $\mathbf{X}_{-i}$. Using Kakutani's theorem on equivalence of product measures it is possible to characterize precisely when

Z can be recovered, but the abstract characterization is not particularly operational. In Appendix B we analyze the characterization for the Gaussian mixture model, where X_i given $Z = z$ has a $\mathcal{N}(\mu_i(z), \sigma_i^2(z))$-distribution. This leads to Proposition 5 and Corollary 1 in Appendix B, which gives necessary and sufficient conditions for recovery in the Gaussian mixture model.

3.2. **Bounding estimation error due to using substitutes.** In this section we derive an upper bound on the estimation error, which is due to using substitutes, cf. the decomposition (13). To this end, we consider the (partly hypothetical) observations $(x_{i,1}, \hat{z}_1, z_1, y_1), \dots (x_{i,n}, \hat{z}_n, z_n, y_n)$, which include the otherwise unobserved z_k-s as well as their observed substitutes, the $\hat{z}_k$-s. We let $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,n})^T \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, and $\|\mathbf{x}_i\|_2$ and $\|\mathbf{y}\|_2$ denote the 2-norms of $\mathbf{x}_i$ and $\mathbf{y}$, respectively. We also let

$$n(z) = \sum_{k=1}^{n} 1(z_k = z) \qquad \text{and} \qquad \hat{n}(z) = \sum_{k=1}^{n} 1(\hat{z}_k = z)$$

for $z \in E = \{1, \dots, K\}$, and

$$n_{\min} = \min\{n(1), \dots, n(K), \hat{n}(1), \dots, \hat{n}(K)\}.$$

Furthermore,

$$\overline{\mu}_i(z) = \frac{1}{n(z)} \sum_{k: z_k = z} x_{i,k},$$

and we define the following three quantities

$$\alpha = \frac{n_{\min}}{n} \tag{16}$$

$$\delta = \frac{1}{n} \sum_{k=1}^{n} 1(\hat{z}_k \neq z_k) \tag{17}$$

$$\rho = \frac{\min\left\{\sum_{k=1}^{n} (x_{i,k} - \overline{\mu}_i(z_k))^2, \sum_{k=1}^{n} (x_{i,k} - \hat{\mu}_i(\hat{z}_k))^2\right\}}{\|\mathbf{x}_i\|_2^2}. \tag{18}$$

Theorem 1. *Let α, δ and ρ be given by* (16), (17) *and* (18). *If $\alpha, \rho > 0$ then*

$$|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \leqslant \frac{2\sqrt{2}}{\rho^2} \sqrt{\frac{\delta}{\alpha}} \frac{\|\mathbf{y}\|_2}{\|\mathbf{x}_i\|_2}. \tag{19}$$

The proof of Theorem 1 is given in Appendix A.2. Appealing to the Law of Large Numbers, the quantities in the upper bound (19) can be interpreted as follows:

- The ratio $\|\mathbf{y}\|_2 / \|\mathbf{x}_i\|_2$ is approximately a fixed and finite constant (unless X_i is constantly zero) depending on the marginal distributions of X_i and Y only.
- The fraction α is approximately

$$\min_{z \in E} \left\{\min\{\mathbb{P}(Z = z), \mathbb{P}(\hat{Z} = z)\}\right\}, \tag{20}$$

  which is strictly positive by Assumption 4(4) (unless recovery is working poorly).

- The quantity ρ is a standardized measure of the residual variation of the $x_{i,k}$-s within the groups defined by the z_k-s or the $\hat{z}_k$-s. It is approximately equal to the constant
$$\frac{\min\left\{\mathbb{E}\left[\mathrm{Var}\left[X_i \mid Z\right]\right], \mathbb{E}\left[\mathrm{Var}\left[X_i \mid \hat{Z}\right]\right]\right\}}{E(X_i^2)},$$
which is strictly positive if the probabilities in (20) are strictly positive and not all of the conditional variances are 0.
- The fraction δ is the relative mislabeling frequency of the substitutes. It is approximately equal to the mislabeling rate $\mathbb{P}(\hat{Z} \neq Z)$.

The bound (19) tells us that if the mislabeling rate of the substitutes tends to 0, that is, if $\mathbb{P}(\hat{Z} \neq Z) \to 0$, the estimation error tends to 0 roughly like $\sqrt{\mathbb{P}(\hat{Z} \neq Z)}$. This could potentially be achieved by letting $p \to \infty$ and $m \to \infty$. We formalize this statement in Section 3.4.

3.3. **Bounding the mislabeling rate of the substitutes.** In this section we give bounds on the mislabeling rate, $\mathbb{P}(\hat{Z} \neq Z)$, with the ultimate purpose of controlling the magnitude of δ in the bound (19). Two different approximations are the culprits of mislabeling. First, the computation of $\hat{Z}$ is based on the p variables in $\mathbf{X}_{1:p}$ only, and it is thus an approximation of the full recovery map based on all variables in $\mathbf{X}$. Second, the recovery map is an estimate and thus itself an approximation. The severity of the second approximation is quantified by the following relative errors of the conditional means used for recovery.

Definition 4 (Relative errors, p-separation)**.** For the mixture model given by Assumption 4 let $\boldsymbol{\mu}_{1:p}(z) = (\mu_i(z))_{i=1,\ldots,p} \in \mathbb{R}^p$ for $z \in E$. With $\check{\boldsymbol{\mu}}_{1:p}(z) \in \mathbb{R}^p$ for $z \in E$ any collection of p-vectors, define the relative errors

$$R_{z,v}^{(p)} = \frac{\|\boldsymbol{\mu}_{1:p}(z) - \check{\boldsymbol{\mu}}_{1:p}(z)\|_2}{\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\|_2} \tag{21}$$

for $z, v \in E$, $v \neq z$. Define, moreover, the minimal p-separation as

$$\mathrm{sep}(p) = \min_{z \neq v} \left\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\right\|_2^2. \tag{22}$$

Note that Assumption 4(2) implies that $\mathrm{sep}(p) \to \infty$ for $p \to \infty$. This convergence could be arbitrarily slow. The following definition captures the important case where the separation grows at least linearly in p.

Definition 5 (Strong separation)**.** We say that the mixture model satisfies *strong separation* if there exists an $\varepsilon > 0$ such that $\mathrm{sep}(p) \geqslant \varepsilon p$ eventually.

Strong separation is equivalent to

$$\liminf_{p \to \infty} \frac{\mathrm{sep}(p)}{p} > 0.$$

A sufficient condition for strong separation is that $|\mu_i(z) - \mu_i(v)| \geqslant \varepsilon$ eventually for all $z, v \in E$, $v \neq z$ and some $\varepsilon > 0$. That is, $\liminf_{i \to \infty} |\mu_i(z) - \mu_i(v)| > 0$ for $v \neq z$. When we

have strong separation, then for p large enough

$$\left(R_{z,v}^{(p)}\right)^2 \leqslant \frac{1}{\varepsilon p}\|\boldsymbol{\mu}_{1:p}(z) - \check{\boldsymbol{\mu}}_{1:p}(z)\|_2^2 \leqslant \frac{1}{\varepsilon}\max_{i=1,\ldots,p}\left(\mu_i(z) - \check{\mu}_i(z)\right)^2,$$

and we note that it is conceivable[2] that we can estimate $\boldsymbol{\mu}_{1:p}(z)$ by an estimator, $\check{\boldsymbol{\mu}}_{1:p}(z)$, such that for $m, p \to \infty$ appropriately, $R_{z,v}^{(p)} \xrightarrow{P} 0$.

The following proposition shows that a bound on $R_{z,v}^{(p)}$ is sufficient to ensure that the growth of $\operatorname{sep}(p)$ controls how fast the mislabeling rate diminishes with p. The proposition is stated for a fixed $\check{\boldsymbol{\mu}}$, which means that when $\check{\boldsymbol{\mu}}$ is an estimate, we are effectively assuming it is independent of the template observation $(\mathbf{X}_{1:p}, Z)$ used to compute $\hat{Z}$.

Proposition 4. *Suppose that Assumption 4 holds. Let $\check{\boldsymbol{\mu}}_{1:p}(z) \in \mathbb{R}^p$ for $z \in E$ and let*

$$\hat{Z} = \arg\min_z \|\mathbf{X}_{1:p} - \check{\boldsymbol{\mu}}_{1:p}(z)\|_2.$$

Suppose also that $R_{z,v}^{(p)} \leqslant \frac{1}{10}$ for all $z, v \in E$ with $v \neq z$. Then

$$\mathbb{P}\left(\hat{Z} \neq Z\right) \leqslant \frac{25K\sigma_{\max}^2}{\operatorname{sep}(p)}. \tag{23}$$

If, in addition, the conditional distribution of X_i given $Z = z$ is sub-Gaussian with variance factor $v_{\max}$, independent of i and z, then

$$\mathbb{P}\left(\hat{Z} \neq Z\right) \leqslant K\exp\left(-\frac{\operatorname{sep}(p)}{50v_{\max}}\right) \tag{24}$$

Remark 6. The proof of Proposition 4 is in Appendix A.3. It shows that the specific constants, 25 and 50, appearing in the bounds above hinge on the specific bound, $R_{z,v}^{(p)} \leqslant \frac{1}{10}$, on the relative error. The proof works for any bound strictly smaller than $\frac{1}{4}$. Replacing $\frac{1}{10}$ by a smaller bound on the relative errors decreases the constant, but it will always be larger than 4.

The upshot of Proposition 4 is that if the relative errors, $R_{z,v}^{(p)}$, are sufficiently small then Assumption 4 is sufficient to ensure that $\mathbb{P}\left(\hat{Z} \neq Z\right) \to 0$ for $p \to \infty$. Without additional distributional assumptions the general bound (23) decays slowly with p, and even with strong separation, the bound only gives a rate of $\frac{1}{p}$. With the additional sub-Gaussian assumption, the rate is improved dramatically, and with strong separation it improves to e^{-cp} for some constant $c > 0$. If the X_i-s are bounded, their (conditional) distributions are sub-Gaussian, thus the rate is fast in this special but important case.

3.4. **Asymptotics of the substitute adjustment estimator.** Suppose Z takes values in $E = \{1, \ldots, K\}$ and that $(x_{i,1}, z_1, y_1), \ldots, (x_{i,n}, z_n, y_n)$ are observations of (X_i, Z, Y). Then Assumption 3 ensures that the oracle OLS estimator $\widehat{\beta}_i$ is $\sqrt{n}$-consistent and that

$$\widehat{\beta}_i \overset{\text{as}}{\approx} \mathcal{N}(\beta_i, w_i^2/n).$$

[2]Parametric assumptions, say, and marginal estimators of each $\mu_i(z)$ that, under Assumption 4, are uniformly consistent over $i \in \mathbb{N}$ can be combined with a simple union bound to show the claim, possibly in a suboptimal way, cf. Section 3.5.

There are standard sandwich formulas for the asymptotic variance parameter w_i^2. In this section we combine the bounds from Sections 3.2 and 3.3 to show our main theoretical result; that $\widehat{\beta}_i^{\text{sub}}$ is a consistent and asymptotically normal estimator of β_i for $n, m \to \infty$ if also $p \to \infty$ appropriately.

Assumption 5. The dataset $\mathcal{S}_0$ in Algorithm 3 consists of i.i.d. observations of $\mathbf{X}_{1:p}$, the dataset $\mathcal{S}$ in Algorithm 3 consists of i.i.d. observations of $(\mathbf{X}_{1:p}, Y)$, and $\mathcal{S}$ is independent of $\mathcal{S}_0$.

Theorem 2. *Suppose Assumption 1 holds and $E(Y^2) < \infty$, and consider the mixture model fulfilling Assumption 4. Consider data satisfying Assumption 5 and the estimator $\widehat{\beta}_i^{\text{sub}}$ given by Algorithm 3. Suppose that $n, m, p \to \infty$ such that $\mathbb{P}(R_{z,v}^{(p)} > \frac{1}{10}) \to 0$. Then the following hold:*

(1) The estimation error due to using substitutes tends to 0 in probability, that is,

$$|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \xrightarrow{P} 0,$$

and $\widehat{\beta}_i^{\text{sub}}$ is a consistent estimator of β_i.

(2) If $\frac{\text{sep}(p)}{n} \to \infty$ and $n\mathbb{P}(R_{z,v}^{(p)} > \frac{1}{10}) \to 0$, then $\sqrt{n}|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \xrightarrow{P} 0$.

(3) If X_i conditionally on $Z = z$ is sub-Gaussian, with variance factor independent of i and z, and if $\frac{\text{sep}(p)}{\log(n)} \to \infty$ and $n\mathbb{P}(R_{z,v}^{(p)} > \frac{1}{10}) \to 0$, then $\sqrt{n}|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \xrightarrow{P} 0$.

In addition, in case (2) as well as case (3), $\widehat{\beta}_i^{\text{sub}} \overset{as}{\approx} \mathcal{N}(\beta_i, w_i^2/n)$, where the asymptotic variance parameter w_i^2 is the same as for the oracle estimator $\widehat{\beta}_i$.

Remark 7. The proof of Theorem 2 is in Appendix A.4. As mentioned in Remark 6, the precise value of the constant $\frac{1}{10}$ is not important. It could be replaced by any other constant *strictly smaller* than $\frac{1}{4}$, and the conclusion would be the same.

Remark 8. The general growth condition on p in terms of n in case (2) is bad; even with strong separation we would need $\frac{p}{n} \to \infty$, that is, p should grow faster than n. In the sub-Gaussian case this improves substantially so that p only needs to grow faster than $\log(n)$.

3.5. **Tensor decompositions.** One open question from both a theoretical and practical perspective is how we construct the estimators $\check{\boldsymbol{\mu}}_{1:p}(z)$. We want to ensure consistency for $m, p \to \infty$, expressed as $\mathbb{P}\left(R_{z,v}^{(p)} > \frac{1}{10}\right) \to 0$ in our theoretical results, and that the estimator can be computed efficiently for large m and p. We indicated in Section 3.3 that simple marginal estimators of $\mu_i(z)$ can achieve this, but such estimators may be highly inefficient. In this section we briefly describe two methods based on tensor decompositions (Anandkumar et al. 2014) related to the third order moments of $\mathbf{X}_{1:p}$. Thus to apply such methods we need to additionally assume that the X_i-s have finite third moments.

Introduce first the third order $p \times p \times p$ tensor $G^{(p)}$ as

$$G^{(p)} = \sum_{i=1}^{p} \mathbf{a}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{a}_i \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{a}_i,$$

where $\mathbf{e}_i \in \mathbb{R}^p$ is the standard basis vector with a 1 in the i-th coordinate and 0 elsewhere, and where

$$\mathbf{a}_i = \sum_{z \in E} \mathbb{P}(Z = z)\sigma_i^2(z)\boldsymbol{\mu}_{1:p}(z).$$

In terms of the third order raw moment tensor and $G^{(p)}$ we define the tensor

$$M_3^{(p)} = \mathbb{E}[\mathbf{X}_{1:p} \otimes \mathbf{X}_{1:p} \otimes \mathbf{X}_{1:p}] - G^{(p)}. \tag{25}$$

Letting $\mathcal{I} = \{(i_1, i_2, i_3) \in \{1, \ldots, p\} \mid i_1, i_2, i_3 \text{ all distinct}\}$ denote the set of indices of the tensors with all entries distinct, we see from the definition of $G^{(p)}$ that $G^{(p)}_{i_1,i_2,i_3} = 0$ for $(i_1, i_2, i_3) \in \mathcal{I}$. Thus

$$(M_3^{(p)})_{i_1,i_2,i_3} = \mathbb{E}\left[X_{i_1} X_{i_2} X_{i_3}\right]$$

for $(i_1, i_2, i_3) \in \mathcal{I}$. In the following, $(M_3^{(p)})_{\mathcal{I}}$ denotes the incomplete tensor obtained by restricting the indices of $M_3^{(p)}$ to $\mathcal{I}$.

The key to using the $M_3^{(p)}$-tensor for estimation of the $\mu_i(z)$-s is the following rank-K tensor decomposition,

$$M_3^{(p)} = \sum_{z=1}^{K} \mathbb{P}(Z = z)\boldsymbol{\mu}_{1:p}(z) \otimes \boldsymbol{\mu}_{1:p}(z) \otimes \boldsymbol{\mu}_{1:p}(z); \tag{26}$$

see Theorem 3.3 in (Anandkumar et al. 2014) or the derivations on page 2 in (Guo, Nie & Yang 2022).

Guo, Nie & Yang (2022) propose an algorithm based on incomplete tensor decomposition as follows: Let $(\widehat{M}_3^{(p)})_{\mathcal{I}}$ denote an estimate of the incomplete tensor $(M_3^{(p)})_{\mathcal{I}}$; obtain an approximate rank-K tensor decomposition of the incomplete tensor $(\widehat{M}_3^{(p)})_{\mathcal{I}}$; extract estimates $\check{\boldsymbol{\mu}}_{1:p}(1), \ldots, \check{\boldsymbol{\mu}}_{1:p}(K)$ from this tensor decomposition. Theorem 4.2 in (Guo, Nie & Yang 2022) shows that if the vectors $\boldsymbol{\mu}_{1:p}(1), \ldots, \boldsymbol{\mu}_{1:p}(K)$ satisfy certain regularity assumptions, they are estimated consistently by their algorithm (up to permutation) if $(\widehat{M}_3^{(p)})_{\mathcal{I}}$ is consistent. We note that the regularity assumptions are fulfilled for generic vectors in $\mathbb{R}^p$.

A computational downside of working directly with $M_3^{(p)}$ is that it grows cubically with p. Anandkumar et al. (2014) propose to consider $\widetilde{\mathbf{X}}^{(p)} = \mathbf{W}^T\mathbf{X}_{1:p} \in \mathbb{R}^K$, where $\mathbf{W}$ is a $p \times K$ whitening matrix. The tensor decomposition is then computed for the corresponding $K \times K \times K$ tensor $\widetilde{M}_3$. When $K < p$ is fixed and p grows, this is computationally advantageous. Theorem 5.1 in Anandkumar et al. (2014) shows that, under a generically satisfied non-degeneracy condition, the tensor decomposition of $\widetilde{M}_3$ can be estimated consistently (up to permutation) if $\widetilde{M}_3$ can be estimated consistently.

To use the methodology from Anandkumar et al. (2014) in Algorithm 3, we replace Step 4 by their Algorithm 1 applied to $\widetilde{\mathbf{x}}^{(0,p)} = \mathbf{W}^T\mathbf{x}_{1:p}^{(0)}$. This will estimate the transformed mean vectors $\widetilde{\boldsymbol{\mu}}^{(p)}(z) = \mathbf{W}^T\boldsymbol{\mu}_{1:p}(z) \in \mathbb{R}^K$. Likewise, we replace Step 5 in Algorithm 3 by

$$\hat{z}_k = \arg\min_z \left\| \widetilde{\mathbf{x}}^{(p)} - \check{\widetilde{\boldsymbol{\mu}}}^{(p)}(z) \right\|_2$$

where $\widetilde{\mathbf{x}}^{(p)} = \mathbf{W}^T \mathbf{x}_{1:p}$. The separation and relative errors conditions should then be expressed in terms of the p-dependent K-vectors $\widetilde{\boldsymbol{\mu}}^{(p)}(1), \ldots, \widetilde{\boldsymbol{\mu}}^{(p)}(K) \in \mathbb{R}^K$.

4. Simulation Study

Our analysis in Section 3 shows that Algorithm 3 is capable of consistently estimating the β_i-parameters via substitute adjustment for $n, m, p \to \infty$ appropriately. The purpose of this section is to shed light on the finite sample performance of substitute adjustment via a simulation study.

The X_i-s are simulated according to a mixture model fulfilling Assumption 4, and the outcome model is as in Example 1, which makes $b_x^i(z) = \mathbb{E}[Y \mid X_i = x; Z = z]$ a partially linear model. Throughout, we take $m = n$ and $\mathcal{S}_0 = \mathcal{S}$ in Algorithm 3. The simulations are carried out for different choices of n, p, $\boldsymbol{\beta}$ and $\mu_i(z)$-s, and we report results on both the mislabeling rate of the latent variables and the mean squared error (MSE) of the β_i-estimators.

4.1. Mixture model simulations and recovery of Z.

The mixture model in our simulations is given as follows.

- We set $K = 10$ and fix $p_{\max} = 1000$ and $n_{\max} = 1000$.
- We draw $\mu_i(z)$-s independently and uniformly from $(-1, 1)$ for $z \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, p_{\max}\}$.
- Fixing the $\mu_i(z)$-s and a choice of $\mu_{\text{scale}} \in \{0.75, 1, 1.5\}$, we simulate $n_{\max}$ independent observations of $(\mathbf{X}_{1:p_{\max}}, Z)$, each with the latent variable Z uniformly distributed on $\{1, ..., K\}$, and X_i given $Z = z$ being $\mathcal{N}(\mu_{\text{scale}} \cdot \mu_i(z), 1)$-distributed.

We use the algorithm from Anandkumar et al. (2014), as described in Section 3.5, for recovery. We replicate the simulation outlined above 10 times, and we consider recovery of Z for $p \in \{50, 100, 200, 1000\}$ and $n \in \{50, 100, 200, 500, 1000\}$. For replication $b \in \{1, \ldots, 10\}$ the actual values of the latent variables are denoted $z_{b,k}$. For each combination of n and p the substitutes are denoted $\hat{z}_{b,k}^{(n,p)}$. The mislabeling rate for fixed p and n is estimated as

$$\delta^{(n,p)} = \frac{1}{10} \sum_{b=1}^{10} \frac{1}{n} \sum_{k=1}^{n} 1(\hat{z}_{b,k}^{(n,p)} \neq z_{b,k}).$$

Figure 2 shows the estimated mislabeling rates from the simulations. The results demonstrate that for reasonable choices of n and p, the algorithm based on (Anandkumar et al. 2014) is capable of recovering Z quite well.

The theoretical upper bounds of the mislabeling rate in Proposition 4 are monotonely decreasing as functions of $\left\| \boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v) \right\|_2$. These are, in turn, monotonely increasing in p and in μ_{scale}. The results in Figure 2 support that this behavior of the upper bounds carry over to the actual mislabeling rate. Moreover, the rapid decay of the mislabeling rate with μ_{scale} is in accordance with the exponential decay of the upper bound in the sub-Gaussian case.

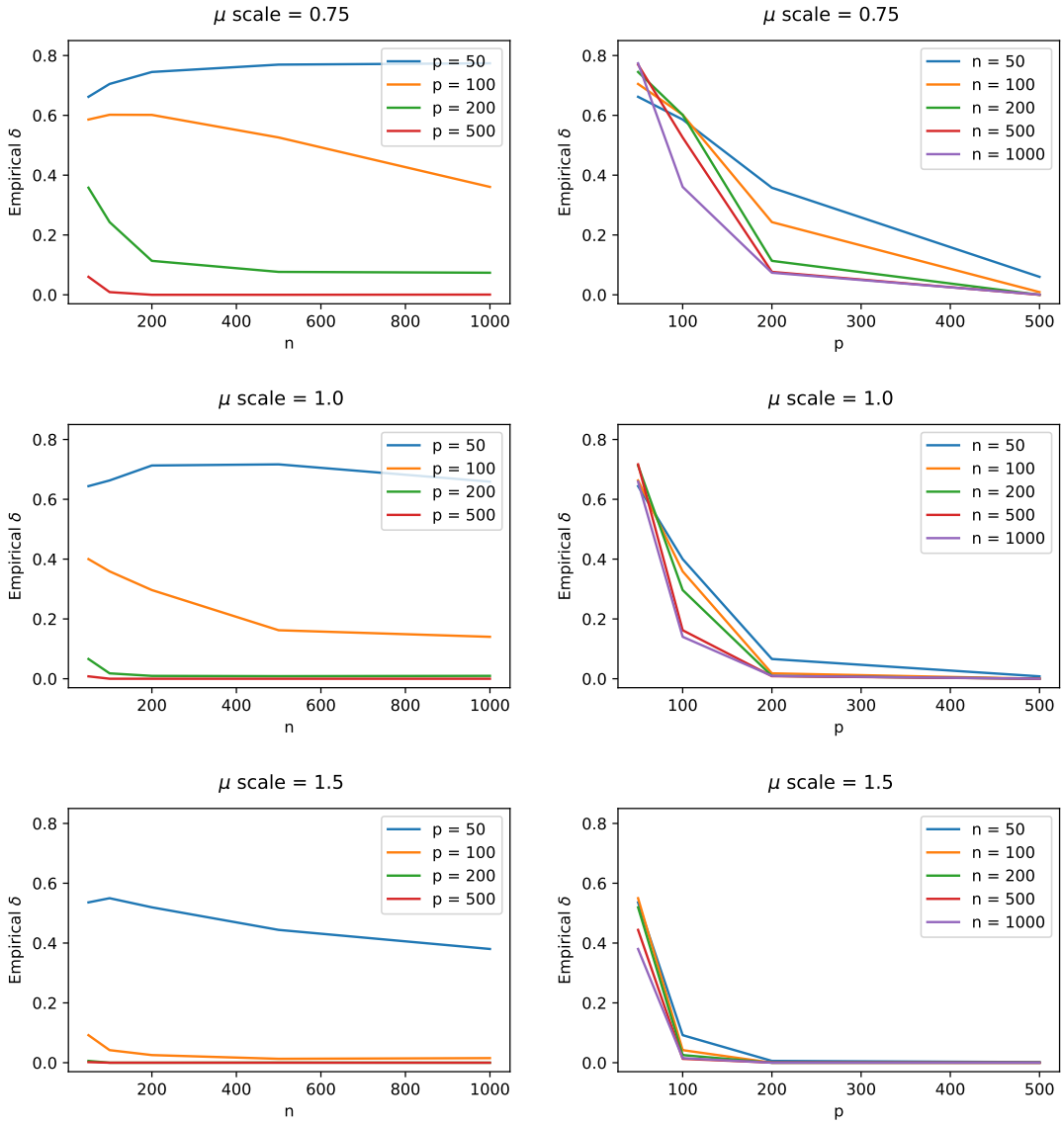


FIGURE 2. Empirical mislabeling rates as a function of $n = m$ and p and for three different separation scales.

4.2. **Outcome model simulation and estimation of β_i.** Given simulated Z-s and X_i-s as described in Section 4.1, we simulate the outcomes as follows.

- Draw β_i independently and uniformly from $(-1, 1)$ for $i = 1, \ldots, p_{\max}$.
- Fix $\gamma_{\text{scale}} \in \{0, 20, 40, 100, 200\}$ and let $\gamma_z = \gamma_{\text{scale}} \cdot z$ for $z \in \{1, \ldots, K\}$.
- With $\varepsilon \sim \mathcal{N}(0, 1)$ simulate $n_{\max}$ independent outcomes as

$$Y = \sum_{i=1}^{p_{\max}} \beta_i X_i + \gamma_Z + \varepsilon.$$

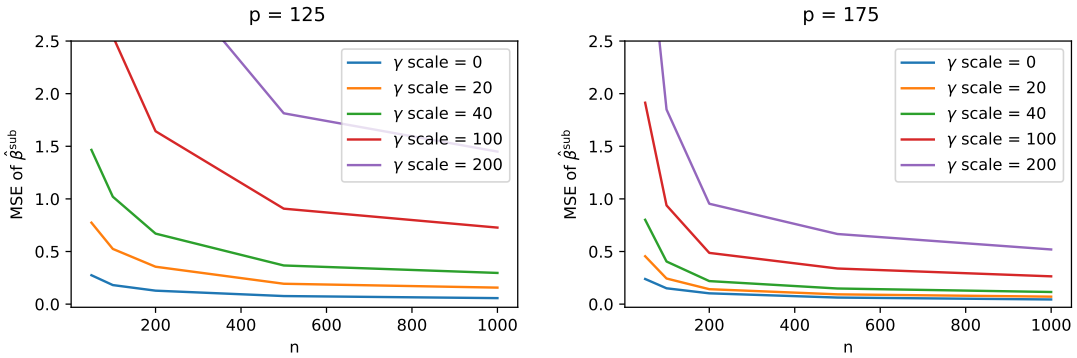


FIGURE 3. Average MSE for substitute adjustment using Algorithm 3 as a function of sample size n and for two different dimensions, a range of the unobserved confounding levels, and with $\mu_{\text{scale}} = 1$.

The simulation parameter γ_{scale} captures a potential effect of unobserved X_i-s for $i > p_{\max}$. We refer to this effect as *unobserved confounding*. For $p < p_{\max}$, adjustment using the naive linear regression model $\sum_{i=1}^{p} \beta_i x_i$ would lead to biased estimates even if $\gamma_{\text{scale}} = 0$, while the naive linear regression model for $p = p_{\max}$ would be correct when $\gamma_{\text{scale}} = 0$. When $\gamma_{\text{scale}} > 0$, adjusting via naive linear regression for all observed X_i-s would still lead to biased estimates due to the unobserved confounding.

We consider the estimation error for $p \in \{125, 175\}$ and $n \in \{50, 100, 200, 500, 1000\}$. Let $\beta_{b,i}$ denote the i-th parameter in the b-th replication, and let $\hat{\beta}_{b,i}^{\text{sub},n,p}$ denote the corresponding estimate from Algorithm 3 for each combination of n and p. The average MSE of $\hat{\boldsymbol{\beta}}_b^{\text{sub},n,p}$ is computed as

$$\text{MSE}^{(n,p)} = \frac{1}{10} \sum_{b=1}^{10} \frac{1}{p} \sum_{i=1}^{p} (\hat{\beta}_{b,i}^{\text{sub},n,p} - \beta_{b,i})^2.$$

Figure 3 shows the MSE for the different combinations of n and p and for different choices of γ_{scale}. Unsurprisingly, the MSE decays with sample size and increases with the magnitude of unobserved confounding. More interestingly, we see a clear decrease with the dimension p indicating that the lower mislabeling rate for larger p translates to a lower MSE as well.

Finally, we compare the results of Algorithm 3 with two other approaches. Letting $\mathbb{X}$ denote the $n \times p$ model matrix for the $x_{i,k}$-s and $\mathbf{y}$ the n-vector of outcomes, the ridge regression estimator is given as

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}^{(n,p)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \min_{\beta_0 \in \mathbb{R}} \|\mathbf{y} - \beta_0 - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

with λ chosen by five-fold cross-validation. The augmented ridge regression estimator is given as

$$\hat{\boldsymbol{\beta}}_{\text{Aug-Ridge}}^{(n,p)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \min_{\boldsymbol{\gamma} \in \mathbb{R}^K} \left\| \mathbf{y} - [\mathbb{X}, \hat{\mathbf{Z}}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

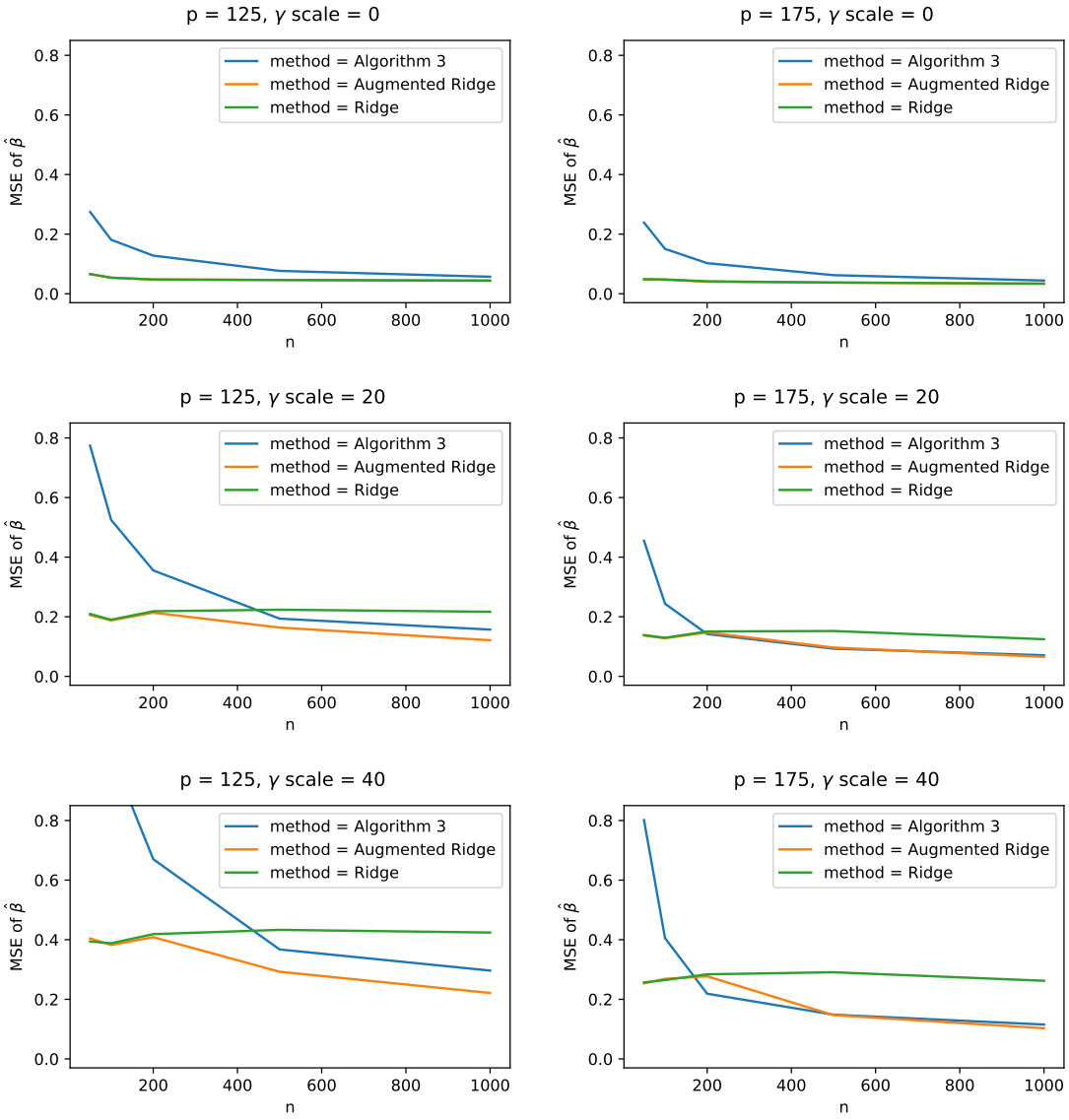


FIGURE 4. Average MSE for substitute adjustment using Algorithm 3 compared to average MSE for the ridge and augmented ridge estimators for two different dimensions, a range of unobserved confounding levels, and with $\mu_{\text{scale}} = 1$.

where $\hat{\mathbf{Z}}$ is the $n \times K$ model matrix of dummy variable encodings of the substitutes. Again, λ is chosen by five-fold cross-validation.

The average MSE is computed for ridge regression and augmented ridge regression just as for substitute adjustment. Figure 4 shows results for $p = 125$ and $p = 175$. These two values of p correspond to asymptotic (as p stays fixed and $n \to \infty$) mislabeling rates δ around 7% and 2%, respectively.

We see that both alternative estimators outperform Algorithm 3 when the sample size is too small to learn Z reliably. However, naive linear regression is biased, and so is ridge regression (even asymptotically), and its performance does not improve as the sample size, n, increases. Substitute adjustment as well as augmented ridge regression adjust for $\hat{Z}$, and their performance improve with n, despite the fact that p is too small to recover Z exactly. When n and the amount of unobserved confounding is sufficiently large, both of these estimators outperform ridge regression. Note that it is unsurprising that the augmented ridge estimator performs similarly to Algorithm 3 for large sample sizes, because after adjusting for the substitutes, the $x_{i,k}$-residuals are roughly orthogonal if the substitutes give accurate recovery, and a joint regression will give estimates similar to those of the marginal regressions.

We made a couple of observations (data not shown) during the simulation study. We experimented with changing the mixture distributions to other sub-Gaussian distributions as well as to the Laplace distribution and got similar results as shown here using the Gaussian distribution. We also implemented sample splitting, and though Proposition 4 assumes sample splitting, we found that the improved estimation accuracy attained by using all available data for the tensor decomposition outweighs the benefit of sample splitting in the recovery stage.

In conclusion, our simulations show that for reasonable finite n and p, it is possible to recover the latent variables sufficiently well for substitute adjustment to be a better alternative than naive linear or ridge regression in settings where the unobserved confounding is sufficiently large.

5. Discussion

We break the discussion into three parts. In the first part we revisit the discussion about the causal interpretation of the target parameters χ^i_x treated in this paper. In the second part we discuss substitute adjustment as a method for estimation of these parameters as well as the assumption-lean parameters β_i. In the third part we discuss possible extensions of our results

5.1. Causal interpretations.

The main causal question is whether a contrast of the form $\chi^i_x - \chi^i_{x_0}$ has a causal interpretation as an average treatment effect. The framework in (Wang & Blei 2019) and the subsequent criticisms by D'Amour (2019) and Ogburn et al. (2020) are based on the X_i-s all being causes of Y, and on the possibility of unobserved confounding. Notably, the latent variable Z to be recovered is not equal to an unobserved confounder, but Wang & Blei (2019) argue that using the deconfounder allows us to weaken the assumption of "no unmeasured confounding" to "no unmeasured single-cause confounding". The assumptions made in (Wang & Blei 2019) did not fully justify this claim, and we found it difficult to understand precisely what the causal assumptions related to Z were.

Mathematically precise assumptions that allow for identification of causal parameters from a finite number of causes, $X_1, \ldots, X_p$, via deconfounding are stated as Assumptions 1 and 2 in (Wang & Blei 2020). We find these assumptions regarding recovery of Z (also termed "pinpointing" in the context of the deconfounder) for finite p implausible. Moreover, the entire framework of the deconfounder rests on the causal assumption

of "weak unconfoundedness" in Assumption 1 and Theorem 1 of (Wang & Blei 2020), which might be needed for a causal interpretation but is unnecessary for the deconfounder algorithm to estimate a meaningful target parameter.

We find it beneficial to disentangle the causal interpretation from the definition of the target parameter. By defining the target parameter entirely in terms of the observational distribution of observed (or, at least, observable) variables, we can discuss the properties of the statistical method of substitute adjustment without making causal claims. We have shown that substitute adjustment under our Assumption 2 on the latent variable model targets the adjusted mean irrespectively of any unobserved confounding. Grimmer et al. (2023) present a similar view. The contrast $\chi^i_x - \chi^i_{x_0}$ might have a causal interpretation in specific applications, but substitute adjustment as a statistical method does not rely on such an interpretation or assumptions needed to justify such an interpretation. In any specific application with multiple causes and potential unobserved confounding, substitute adjustment might be a useful method for deconfounding, but depending on the context and the causal assumptions we are willing to make, other methods could be preferable (Miao et al. 2023).

5.2. **Substitute adjustment: interpretation, merits and deficits.** We define the target parameter as an adjusted mean when adjusting for an *infinite* number of variables. Clearly, this is a mathematical idealization of adjusting for a large number of variables, but it also has some important technical consequences. First, the recovery Assumption 2(2) is a more plausible modelling assumption than recovery from a finite number of variables. Second, it gives a clear qualitative difference between the adjusted mean of one (or any finite number of) variables and regression on all variables. Third, the natural requirement in Assumption 2(2) that Z can be recovered from $\mathbf{X}_{-i}$ for any i replaces the minimality of a "multi-cause separator" from (Wang & Blei 2020). Our assumption is that $\sigma(Z)$ is sufficiently minimal in a very explicit way, which ensures that Z does not contain information unique to any single X_i.

Grimmer et al. (2023) come to a similar conclusion as we do: that the target parameter of substitute adjustment (and the deconfounder) is the adjusted mean χ^i_x, where you adjust for an infinite number of variables. They argue forcefully that substitute adjustment, using a finite number p of variables, does not have an advantage over naive regression, that is, over estimating the regression function $\mathbb{E}\left[Y \mid X_1 = x_1, \ldots, X_p = x_p\right]$ directly. With $i = 1$, say, they argue that substitute adjustment is effectively assuming a partially linear, semiparametric regression model

$$\mathbb{E}\left[Y \mid X_1 = x_1, \ldots, X_p = x_p\right] = \beta_0 + \beta_1 x_1 + h(x_2, \ldots, x_p),$$

with the specific constraint that $h(x_2, \ldots, x_p) = g(\hat{z}) = g(f^{(p)}(x_2, \ldots, x_p))$. We agree with their analysis and conclusion; substitute adjustment is implicitly a way of making assumptions about h. It is also a way to leverage those assumptions, either by shrinking the bias compared to directly estimating a misspecified (linear, say) h, or by improving efficiency over methods that use a too flexible model of h. We believe there is room for further studies of such bias and efficiency tradeoffs.

We also believe that there are two potential benefits of substitute adjustment, which are not brought forward by Grimmer et al. (2023). First, the latent variable model can be estimated without access to outcome observations. This means that the inner part of

$h = g \circ f^{(p)}$ could, potentially, be estimated very accurately on the basis of a large sample $\mathcal{S}_0$ in cases where it would be difficult to estimate the composed map h accurately from $\mathcal{S}$ alone. Second, when p is very large, e.g., in the millions, but Z is low-dimensional, there can be huge computational advantages to running p small parallel regressions compared to just one naive linear regression of Y on all of $\mathbf{X}_{1:p}$, let alone p naive partially linear regressions.

5.3. **Possible extensions.** We believe that our error bound in Theorem 1 is an interesting result, which in a precise way bounds the error of an OLS estimator in terms of errors in the regressors. This result is closely related to the classical literature on errors-in-variables models (or measurement error models) (Durbin 1954, Cochran 1968, Schennach 2016), though this literature focuses on methods for bias correction when the errors are non-vanishing. We see two possible extensions of our result. For one, Theorem 1 could easily be generalized to $E = \mathbb{R}^d$. In addition, it might be possible to apply the bias correction techniques developed for errors-in-variables to improve the finite sample properties of the substitute adjustment estimator.

Our analysis of the recovery error could also be extended. The concentration inequalities in Section 3.3 are unsurprising, but developed to match our specific needs for a high-dimensional analysis with as few assumptions as possible. For more refined results on finite mixture estimation see, e.g., (Heinrich & Kahn 2018), and see (Ndaoud 2022) for optimal recovery when $K = 2$ and the mixture distributions are Gaussian. In cases where the mixture distributions are Gaussian, it is also plausible that specialized algorithms such as (Kalai et al. 2012, Gandhi & Borns-Weil 2016) are more efficient than the methods we consider based on conditional means only.

One general concern with substitute adjustment is model misspecification. We have done our analysis with minimal distributional assumptions, but there are, of course, two fundamental assumptions: the assumption of conditional independence of the X_i-s given the latent variable Z, and the assumption that Z takes values in a finite set of size K. An important extension of our results is to study robustness to violations of these two fundamental assumptions. We have also not considered estimation of K, and it would likewise be relevant to understand how that affects the substitute adjustment estimator.

Acknowledgments

We thank Alexander Mangulad Christgau for helpful input. JA and NRH were supported by a research grant (NNF20OC0062897) from Novo Nordisk Fonden. JA also received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 801199.

Appendix A. Proofs and auxiliary results

A.1. **Proofs of results in Section 2.1.**

Proof of Proposition 1. Since X_i as well as $\mathbf{X}_{-i}$ take values in Borel spaces, there exists a regular conditional distribution given $Z = z$ of each (Kallenberg 2021, Theorem 8.5). These are denoted P_z^i and P_z^{-i}, respectively. Moreover, Assumption 2(2) and the Doob-Dynkin lemma (Kallenberg 2021, Lemma 1.14) imply that for each $i \in \mathbb{N}$ there is a

measurable map $f_i : \mathbb{R}^{\mathbb{N}} \to E$ such that $Z = f_i(\mathbf{X}_{-i})$. This implies that $P^{-i}(B) = \int P_z^{-i}(B) P^Z(\mathrm{d}z)$ for $B \subseteq \mathbb{R}^{\mathbb{N}}$ measurable.

Since $Z = f_i(\mathbf{X}_{-i})$ it holds that $f_i(P^{-i}) = P^Z$, and furthermore that $P_z^{-i}(f_i^{-1}(\{z\})) = 1$. Assumption 2(1) implies that X_i and $\mathbf{X}_{-i}$ are conditionally independent given Z, thus for $A, C \subseteq \mathbb{R}$ and $B \subseteq E$ measurable sets and $\tilde{B} = f_i^{-1}(B) \subseteq \mathbb{R}^{\mathbb{N}}$,

$$\begin{aligned}
\mathbb{P}(X_i \in A, Z \in B, Y \in C) &= \mathbb{P}(X_i \in A, \mathbf{X}_{-i} \in \tilde{B}, Y \in C) \\
&= \int 1_A(x) 1_{\tilde{B}}(\mathbf{x}) P^i_{x,\mathbf{x}}(C) P(\mathrm{d}x, \mathrm{d}\mathbf{x}) \\
&= \int 1_A(x) 1_{\tilde{B}}(\mathbf{x})\, P^i_{x,\mathbf{x}}(C) \int P^i_z \otimes P_z^{-i}(\mathrm{d}x, \mathrm{d}\mathbf{x}) P^Z(\mathrm{d}z) \\
&= \iiint 1_A(x) 1_{\tilde{B}}(\mathbf{x})\, P^i_{x,\mathbf{x}}(C) P^i_z(\mathrm{d}x) P_z^{-i}(\mathrm{d}\mathbf{x}) P^Z(\mathrm{d}z) \\
&= \iiint 1_A(x) 1_B(z) \int P^i_{x,\mathbf{x}}(C) P_z^{-i}(\mathrm{d}\mathbf{x}) P^i_z(\mathrm{d}x) P^Z(\mathrm{d}z) \\
&= \iint 1_A(x) 1_B(z)\, Q^i_{x,z}(C) P^i_z(\mathrm{d}x) P^Z(\mathrm{d}z).
\end{aligned}$$

Hence $Q^i_{x,z}$ is a regular conditional distribution of Y given $(X_i, Z) = (x, z)$.

We finally find that

$$\begin{aligned}
\chi^i_x &= \iint y\, P^i_{x,\mathbf{x}}(\mathrm{d}y) P^{-i}(\mathrm{d}\mathbf{x}) \\
&= \iiint y\, P^i_{x,\mathbf{x}}(\mathrm{d}y) P_z^{-i}(\mathrm{d}\mathbf{x}) P^Z(\mathrm{d}z) \\
&= \iint y \int P^i_{x,\mathbf{x}}(\mathrm{d}y) P_z^{-i}(\mathrm{d}\mathbf{x}) P^Z(\mathrm{d}z) \\
&= \iint y\, Q^i_{z,\mathbf{x}}(\mathrm{d}y) P^Z(\mathrm{d}z).
\end{aligned}$$

□

Proof of Proposition 2. We find that

$$\begin{aligned}
\mathrm{Cov}\left[X_i, Y \mid Z\right] &= \mathbb{E}\left[(X_i - \mathbb{E}[X_i \mid Z])Y \mid Z\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(X_i - \mathbb{E}[X_i \mid Z])Y \mid X_i, Z\right] \mid Z\right] \\
&= \mathbb{E}\left[(X_i - \mathbb{E}[X_i \mid Z])\mathbb{E}\left[Y \mid X_i, Z\right] \mid Z\right] \\
&= \mathbb{E}\left[(X_i - \mathbb{E}[X_i \mid Z]) b^i_{X_i}(Z) \mid Z\right] \\
&= \mathrm{Cov}\left[X_i, b^i_{X_i}(Z) \mid Z\right],
\end{aligned}$$

which shows (9). From this representation, if $b^i_x(z) = b^i(z)$ does not depend on x, $b^i(Z)$ is $\sigma(Z)$-measurable and $\mathrm{Cov}\left[X_i, b^i(Z) \mid Z\right] = 0$, whence $\beta_i = 0$.

If $b_x^i(z) = \beta_i'(z)x + \eta_{-i}(z)$,

$$\operatorname{Cov}\left[X_i, b^i_{X_i}(Z) \mid Z\right] = \operatorname{Cov}\left[X_i, \beta_i'(Z)X_i + \eta_{-i}(Z) \mid Z\right] = \beta_i'(Z)\operatorname{Var}\left[X_i \mid Z\right],$$

and (10) follows. □

A.2. **Auxiliary results related to Section 3.2 and proof of Theorem 1.** Let $\mathbf{Z}$ denote the $n \times K$ matrix of dummy variable encodings of the z_k-s, and let $\hat{\mathbf{Z}}$ denote the similar matrix for the substitutes $\hat{z}_k$-s. With $P_{\mathbf{Z}}$ and $P_{\hat{\mathbf{Z}}}$ the orthogonal projections onto the column spaces of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, respectively, we can write the estimator from Algorithm 3 as

$$\widehat{\beta}_i^{\text{sub}} = \frac{\langle \mathbf{x}_i - P_{\hat{\mathbf{Z}}}\mathbf{x}_i, \mathbf{y} - P_{\hat{\mathbf{Z}}}\mathbf{y}\rangle}{\|\mathbf{x}_i - P_{\hat{\mathbf{Z}}}\mathbf{x}_i\|_2^2}. \tag{27}$$

Here $\mathbf{x}_i, \mathbf{y} \in \mathbb{R}^n$ denote the n-vectors of $x_{i,k}$-s and y_k-s, respectively, and $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^n$, so that, e.g., $\|\mathbf{y}\|_2^2 = \langle \mathbf{y}, \mathbf{y}\rangle$. The estimator, had we observed the latent variables, is similarly given as

$$\hat{\beta}_i = \frac{\langle \mathbf{x}_i - P_{\mathbf{Z}}\mathbf{x}_i, \mathbf{y} - P_{\mathbf{Z}}\mathbf{y}\rangle}{\|\mathbf{x}_i - P_{\mathbf{Z}}\mathbf{x}_i\|_2^2}. \tag{28}$$

The proof of Theorem 1 is based on the following bound on the difference between the projection matrices.

Lemma 1. *Let α and δ be as defined by* (16) *and* (17)*. If $\alpha > 0$ it holds that*

$$\|P_{\mathbf{Z}} - P_{\hat{\mathbf{Z}}}\|_2 \leqslant \sqrt{\frac{2\delta}{\alpha}}, \tag{29}$$

where $\|\cdot\|_2$ above denotes the operator 2-norm also known as the spectral norm.

Proof. When $\alpha > 0$, the matrices $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ have full rank K. Let $\mathbf{Z}^+ = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ and $\hat{\mathbf{Z}}^+ = (\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}^T$ denote the Moore-Penrose inverses of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, respectively. Then $P_{\mathbf{Z}} = \mathbf{Z}\mathbf{Z}^+$ and $P_{\hat{\mathbf{Z}}} = \hat{\mathbf{Z}}\hat{\mathbf{Z}}^+$. By Theorems 2.3 and 2.4 in (Stewart 1977),

$$\|P_{\mathbf{Z}} - P_{\hat{\mathbf{Z}}}\|_2 \leqslant \min\left\{\|\mathbf{Z}^+\|_2, \|\hat{\mathbf{Z}}^+\|_2\right\}\ \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2.$$

The operator 2-norm $\|\mathbf{Z}^+\|_2$ is the square root of the largest eigenvalue of

$$(\mathbf{Z}^T\mathbf{Z})^{-1} = \begin{pmatrix} n(1)^{-1} & 0 & \dots & 0 \\ 0 & n(2)^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n(K)^{-1} \end{pmatrix}.$$

Whence $\|\mathbf{Z}^+\|_2 \leqslant (n_{\min})^{-1/2} = (\alpha n)^{-1/2}$. The same bound is obtained for $\|\hat{\mathbf{Z}}^+\|_2$, which gives

$$\|P_{\mathbf{Z}} - P_{\hat{\mathbf{Z}}}\|_2 \leqslant \frac{1}{\sqrt{\alpha n}}\ \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2.$$

We also have that

$$\|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \leqslant \|\mathbf{Z} - \hat{\mathbf{Z}}\|_F^2 = \sum_{k=1}^{n}\sum_{i=1}^{p}(\mathbf{Z}_{k,i} - \hat{\mathbf{Z}}_{k,i})^2 = 2\delta n,$$

because $\sum_{i=1}^{p}(\mathbf{Z}_{k,i} - \hat{\mathbf{Z}}_{k,i})^2 = 2$ precisely for those k with $\hat{z}_k \neq z_k$ and 0 otherwise. Combining the inequalities gives (29). □

Before proceeding with the proof of Theorem 1, note that

$$\sum_{k=1}^{n}(x_{i,k}-\overline{\mu}_i(z_k))^2 = \|\mathbf{x}_i - P_{\mathbf{Z}}\mathbf{x}_i\|_2^2 = \|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2 \leqslant \|\mathbf{x}_i\|_2^2$$

since $(I-P_{\mathbf{Z}})$ is a projection. Similarly, $\sum_{k=1}^{n}(x_{i,k}-\hat{\mu}_i(\hat{z}_k))^2 = \|\mathbf{x}_i - P_{\hat{\mathbf{Z}}}\mathbf{x}_i\|_2^2 \leqslant \|\mathbf{x}\|_2^2$, thus

$$\rho = \frac{\min\left\{\|\mathbf{x}_i - P_{\mathbf{Z}}\mathbf{x}_i\|_2^2, \|\mathbf{x}_i - P_{\hat{\mathbf{Z}}}\mathbf{x}_i\|_2^2\right\}}{\|\mathbf{x}_i\|_2^2} \leqslant 1.$$

Proof of Theorem 1. First note that since $I - P_{\hat{\mathbf{Z}}}$ is an orthogonal projection,

$$\langle \mathbf{x}_i - P_{\hat{\mathbf{Z}}}\mathbf{x}_i, \mathbf{y} - P_{\hat{\mathbf{Z}}}\mathbf{y}\rangle = \langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle$$

and similarly for the other inner product in (28). Moreover,

$$\langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle - \langle \mathbf{x}_i, (I-P_{\mathbf{Z}})\mathbf{y}\rangle = \langle \mathbf{x}_i, (P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle$$

and

$$\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2 - \|(I-P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2 = \|(P_{\hat{\mathbf{Z}}}-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2.$$

We find that

$$\begin{aligned}
\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i &= \frac{\langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle}{\|(I-P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2} - \frac{\langle \mathbf{x}_i, (I-P_{\mathbf{Z}})\mathbf{y}\rangle}{\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2} \\
&= \langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle \left(\frac{1}{\|(I-P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2} - \frac{1}{\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2}\right) \\
&\qquad + \frac{\langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle - \langle \mathbf{x}_i, (I-P_{\mathbf{Z}})\mathbf{y}\rangle}{\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2} \\
&= \langle \mathbf{x}_i, (I-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle \left(\frac{\|(P_{\hat{\mathbf{Z}}}-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2}{\|(I-P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2}\right) \\
&\qquad + \frac{\langle \mathbf{x}_i, (P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}})\mathbf{y}\rangle}{\|(I-P_{\mathbf{Z}})\mathbf{x}_i\|_2^2}.
\end{aligned}$$

This gives the following inequality, using that $\rho \leqslant 1$,

$$\begin{aligned}
|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| &\leqslant \frac{\|P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}}\|_2\|\mathbf{x}_i\|_2^3\|\mathbf{y}\|_2}{\rho^2\|\mathbf{x}_i\|_2^4} + \frac{\|P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}}\|_2\|\mathbf{x}_i\|_2\|\mathbf{y}\|_2}{\rho\|\mathbf{x}_i\|_2^2} \\
&= \left(\frac{1}{\rho^2}+\frac{1}{\rho}\right)\|P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}}\|_2\frac{\|\mathbf{y}\|_2}{\|\mathbf{x}_i\|_2} \\
&\leqslant \frac{2}{\rho^2}\|P_{\mathbf{Z}}-P_{\hat{\mathbf{Z}}}\|_2\frac{\|\mathbf{y}\|_2}{\|\mathbf{x}_i\|_2}.
\end{aligned}$$

Combining this inequality with (29) gives (19). □

A.3. **Auxiliary concentration inequalities. Proofs of Propositions 3 and 4.**

Lemma 2. *Suppose that Assumption 4 holds. Let* $\check{\boldsymbol{\mu}}_{1:p}(z) \in \mathbb{R}^p$ *for* $z \in E$ *and let* $\hat{Z} = \arg\min_z \|\mathbf{X}_{1:p} - \check{\boldsymbol{\mu}}_{1:p}(z)\|_2$. *Suppose that* $R_{z,v}^{(p)} \leqslant \frac{1}{10}$ *for all* $z, v \in E$ *with* $v \neq z$ *then*

$$\mathbb{P}(\hat{Z} = v \mid Z = z) \leqslant \frac{25\sigma_{\max}^2}{\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\|_2^2}. \tag{30}$$

Proof. Since p is fixed throughout the proof, we simplify the notation by dropping the $1{:}p$ subscript and use, e.g., $\mathbf{X}$ and $\boldsymbol{\mu}$ to denote the $\mathbb{R}^p$-vectors $\mathbf{X}_{1:p}$ and $\boldsymbol{\mu}_{1:p}$, respectively.

Fix also $z, v \in E$ with $v \neq z$ and observe first that

$$\begin{aligned}(\hat{Z} = v) &\subseteq (\|\mathbf{X} - \check{\boldsymbol{\mu}}(v)\|_2 < \|\mathbf{X} - \check{\boldsymbol{\mu}}(z)\|_2) \\ &= \left(\langle \mathbf{X} - \check{\boldsymbol{\mu}}(z), \check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\rangle < -\tfrac{1}{2}\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2^2\right) \\ &= \Big(\langle \mathbf{X} - \boldsymbol{\mu}(z), \check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\rangle < \\ &\qquad - \left(\tfrac{1}{2}\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2^2 + \langle \boldsymbol{\mu}(z) - \check{\boldsymbol{\mu}}(z), \check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\rangle\right)\Big).\end{aligned}$$

The objective is to bound the probability of the event above using Chebyshev's inequality. To this end, we first use the Cauchy-Schwarz inequality to get

$$\begin{aligned}&\tfrac{1}{2}\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2^2 + \langle \boldsymbol{\mu}(z) - \check{\boldsymbol{\mu}}(z), \check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\rangle \\ &\qquad \geqslant \tfrac{1}{2}\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2^2 - \|\boldsymbol{\mu}(z) - \check{\boldsymbol{\mu}}(z)\|_2\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2 \\ &\qquad = \|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^2 \left(\tfrac{1}{2}B_{z,v}^2 - R_{z,v}^{(p)} B_{z,v}\right),\end{aligned}$$

where

$$B_{z,v} = \frac{\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2}{\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2}.$$

The triangle and reverse triangle inequality give that

$$\begin{aligned}\|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2 &\leqslant \|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2 + \|\check{\boldsymbol{\mu}}(z) - \boldsymbol{\mu}(z)\|_2 + \|\boldsymbol{\mu}(v) - \check{\boldsymbol{\mu}}(v)\|_2 \\ \|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2 &\geqslant \Big|\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2 - \|\boldsymbol{\mu}(z) - \check{\boldsymbol{\mu}}(z)\|_2 - \|\boldsymbol{\mu}(v) - \check{\boldsymbol{\mu}}(v)\|_2\Big|,\end{aligned}$$

and dividing by $\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2$ combined with the bound $\frac{1}{10}$ on the relative errors yield

$$\begin{aligned}B_{z,v} &\leqslant 1 + R_{z,v}^{(p)} + R_{v,z}^{(p)} \leqslant \frac{6}{5}, \\ B_{z,v} &\geqslant \left|1 - R_{z,v}^{(p)} - R_{v,z}^{(p)}\right| \geqslant \frac{4}{5}.\end{aligned}$$

This gives

$$\tfrac{1}{2}B_{z,v}^2 - R_{z,v}^{(p)} B_{z,v} \geqslant \tfrac{1}{2}B_{z,v}^2 - \tfrac{1}{10}B_{z,v} \geqslant \tfrac{6}{25}$$

since the function $b \mapsto b^2 - \frac{2}{10}b$ is increasing for $b \geqslant \frac{4}{5}$.

Introducing the variables $W_i = (X_i - \mu_i(z))(\check{\mu}_i(z) - \check{\mu}_i(v))$ we conclude that

$$(\hat{Z} = v) \subseteq \left(\sum_{i=1}^{p} W_i < -\tfrac{6}{25}\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^2\right). \tag{31}$$

Note that $E[W_i \mid Z = z] = 0$ and $\mathrm{Var}[W_i \mid Z = z] = (\check{\mu}_i(z) - \check{\mu}_i(v))^2\sigma_i^2(z)$, and by Assumption 4, the W_i-s are conditionally independent given $Z = z$, so Chebyshev's

inequality gives that

$$\begin{aligned}
\mathbb{P}(\hat{Z} = v \mid Z = z) &\leqslant \mathbb{P}\left(\sum_{i=1}^{p} W_i < -\tfrac{6}{25}\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^2 \,\middle|\, Z = z\right) \\
&\leqslant \left(\frac{25}{6}\right)^2 \frac{\sum_{i=1}^{p} (\check{\mu}_i(z) - \check{\mu}_i(v))^2 \sigma_i^2(z)}{\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^4} \\
&\leqslant \left(\frac{25}{6}\right)^2 \frac{\sigma_{\max}^2 \|\check{\boldsymbol{\mu}}(z) - \check{\boldsymbol{\mu}}(v)\|_2^2}{\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_4^2} \\
&\leqslant \left(\frac{25}{6}\right)^2 B_{z,v}^2 \frac{\sigma_{\max}^2}{\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^2} \\
&\leqslant \frac{25\sigma_{\max}^2}{\|\boldsymbol{\mu}(z) - \boldsymbol{\mu}(v)\|_2^2},
\end{aligned}$$

where we, for the last inequality, used that $B_{z,v}^2 \leqslant \left(\frac{6}{5}\right)^2$. □

Before proceeding to the concentration inequality for sub-Gaussian distributions, we use Lemma 2 to prove Proposition 3.

Proof of Proposition 3. Suppose that $i = 1$ for convenience. We take $\check{\boldsymbol{\mu}}_{1:p}(z) = \boldsymbol{\mu}_{1:p}(z)$ for all $p \in \mathbb{N}$ and $z \in E$ and write $\hat{Z}_p = \arg\min_z \|\mathbf{X}_{2:p} - \boldsymbol{\mu}_{2:p}(z)\|_2$ for the prediction of Z based on the coordinates $2, \ldots, p$. With this oracle choice of $\check{\boldsymbol{\mu}}_{1:p}(z)$, the relative errors are zero, thus the bound (30) holds, and Lemma 2 gives

$$\begin{aligned}
\mathbb{P}\left(\hat{Z}_p \neq Z\right) &= \sum_z \sum_{v \neq z} \mathbb{P}\left(\hat{Z}_p = v, Z = z\right) \\
&= \sum_z \sum_{v \neq z} \mathbb{P}\left(\hat{Z}_p = v \mid Z = z\right) \mathbb{P}\left(Z = z\right) \\
&\leqslant \frac{C}{\min_{z \neq v} \left\|\boldsymbol{\mu}_{2:p}(z) - \boldsymbol{\mu}_{2:p}(v)\right\|_2^2}
\end{aligned}$$

with C a constant independent of p. By (14), $\min_{z \neq v} \left\|\boldsymbol{\mu}_{2:p}(z) - \boldsymbol{\mu}_{2:p}(v)\right\|_2^2 \to \infty$ for $p \to \infty$, and by choosing a subsequence, p_r, we can ensure that $\mathbb{P}\left(\hat{Z}_{p_r} \neq Z\right) \leqslant \frac{1}{r^2}$. Then $\sum_{r=1}^{\infty} \mathbb{P}\left(\hat{Z}_{p_r} \neq Z\right) < \infty$, and by Borel-Cantelli's lemma,

$$\mathbb{P}\left(\hat{Z}_{p_r} \neq Z \text{ infinitely often}\right) = 0.$$

That is, $\mathbb{P}\left(\hat{Z}_{p_r} = Z \text{ eventually}\right) = 1$, which shows that we can recover Z from $(\hat{Z}_{p_r})_{r \in \mathbb{N}}$ and thus from $\mathbf{X}_{-1}$ (with probability 1). Defining

$$Z' = \begin{cases} \lim\limits_{r \to \infty} \hat{Z}_{p_r} & \text{if } \hat{Z}_{p_r} = Z \text{ eventually} \\ 0 & \text{otherwise} \end{cases}$$

we see that $\sigma(Z') \subseteq \sigma(\mathbf{X}_{-1})$ and $Z' = Z$ almost surely. Thus if we replace Z by Z' in Assumption 4 we see that Assumption 2(2) holds. □

Lemma 3. *Consider the same setup as in Lemma 2, that is, Assumption 4 holds and* $R_{z,v}^{(p)} \leqslant \frac{1}{10}$ *for all* $z, v \in E$ *with* $v \neq z$*. Suppose, in addition, that the conditional distribution of* X_i *given* $Z = z$ *is sub-Gaussian with variance factor* $v_{\max}$*, independent of* i *and* z*, then*

$$\mathbb{P}(\hat{Z} = v \mid Z = z) \leqslant \exp\left(-\frac{1}{50 v_{\max}} \|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\|_2^2\right). \tag{32}$$

Proof. Recall that X_i given $Z = z$ being sub-Gaussian with variance factor $v_{\max}$ means that

$$\log \mathbb{E}\left[e^{\lambda(X_i - \mu_i(z))} \,\middle|\, Z = z\right] \leqslant \frac{1}{2}\lambda^2 v_{\max}$$

for $\lambda \in \mathbb{R}$. Consequently, with W_i as in the proof of Lemma 2, and using conditional independence of the X_i-s given $Z = z$,

$$\begin{aligned}
\log \mathbb{E}\left[e^{\lambda \sum_{i=1}^p W_i} \,\middle|\, Z = z\right] &= \sum_{i=1}^p \log \mathbb{E}\left[e^{\lambda(\check{\mu}_i(z) - \check{\mu}_i(v))(X_i - \mu_i(z))} \,\middle|\, Z = z\right] \\
&\leqslant \frac{1}{2}\lambda^2 v_{\max} \sum_{i=1}^p (\check{\mu}_i(z) - \check{\mu}_i(v))^2 \\
&= \frac{1}{2}\lambda^2 v_{\max} \|\check{\boldsymbol{\mu}}_{1:p}(z) - \check{\boldsymbol{\mu}}_{1:p}(v)\|_2^2.
\end{aligned}$$

Using (31) in combination with the Chernoff bound gives

$$\begin{aligned}
\mathbb{P}(\hat{Z} = v \mid Z = z) &\leqslant \mathbb{P}\left(\sum_{i=1}^p W_i < -\tfrac{6}{25}\left\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\right\|_2^2 \,\middle|\, Z = z\right) \\
&\leqslant \exp\left(-\left(\frac{6}{25}\right)^2 \frac{\left\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\right\|_2^4}{2 v_{\max} \|\check{\boldsymbol{\mu}}_{1:p}(z) - \check{\boldsymbol{\mu}}_{1:p}(v)\|_2^2}\right) \\
&= \exp\left(-\frac{1}{2 v_{\max}}\left(\frac{6}{25}\right)^2 B_{z,v}^{-2}\left\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\right\|_2^2\right) \\
&\leqslant \exp\left(-\frac{1}{50 v_{\max}}\left\|\boldsymbol{\mu}_{1:p}(z) - \boldsymbol{\mu}_{1:p}(v)\right\|_2^2\right),
\end{aligned}$$

where we, as in the proof of Lemma 2, have used that the bound on the relative error implies that $B_{z,v} \leqslant \frac{6}{5}$. □

Proof of Proposition 4. The argument proceeds as in the proof of Proposition 3. We first note that

$$\begin{aligned}
\mathbb{P}\left(\hat{Z} \neq Z\right) &= \sum_z \sum_{v \neq z} \mathbb{P}\left(\hat{Z} = v, Z = z\right) \\
&= \sum_z \sum_{v \neq z} \mathbb{P}\left(\hat{Z} = v \,\middle|\, Z = z\right) \mathbb{P}\left(Z = z\right).
\end{aligned}$$

Lemma 2 then gives

$$\mathbb{P}\left(\hat{Z} \neq Z\right) \leqslant \frac{25 K \sigma_{\max}^2}{\operatorname{sep}(p)}.$$

If the sub-Gaussian assumption holds, Lemma 3 instead gives

$$\mathbb{P}\left(\hat{Z} \neq Z\right) \leqslant K \exp\left(-\frac{\operatorname{sep}(p)}{50 v_{\max}}\right).$$

□

A.4. **Proof of Theorem 2.**

Proof of Theorem 2. Recall that

$$\delta = \frac{1}{n}\sum_{k=1}^{n} 1(\hat{z}_k \neq z_k),$$

hence by Proposition 4

$$\begin{aligned}
\mathbb{E}[\delta] &= \mathbb{P}(\hat{Z}_k \neq Z) \\
&\leqslant \mathbb{P}\left(\hat{Z}_k \neq Z \,\middle|\, \max_{z \neq v} R_{z,v}^{(p)} \leqslant \tfrac{1}{10}\right) + \mathbb{P}\left(\max_{z \neq v} R_{z,v}^{(p)} > \tfrac{1}{10}\right) \\
&\leqslant \frac{25 K \sigma_{\max}^2}{\operatorname{sep}(p)} + K^2 \max_{z \neq v} \mathbb{P}\left(R_{z,v}^{(p)} > \tfrac{1}{10}\right). \qquad (33)
\end{aligned}$$

Both of the terms above tend to 0, thus $\delta \xrightarrow{P} 0$.

Now rewrite the bound (19) as

$$|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \leqslant \sqrt{\delta} \underbrace{\left(\frac{2\sqrt{2}}{\rho^2 \sqrt{\alpha}} \frac{\|\mathbf{y}\|_2}{\|\mathbf{x}_i\|_2}\right)}_{=L_n}$$

From the argument above, $\sqrt{\delta} \xrightarrow{P} 0$. We will show that the second factor, L_n, tends to a constant, L, in probability under the stated assumptions. This will imply that

$$|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \xrightarrow{P} 0,$$

which shows case (1).

Observe first that

$$\|\mathbf{x}_i\|_2^2 = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}^2 \xrightarrow{P} \mathbb{E}[X_i^2] \in (0, \infty)$$

by the Law of Large Numbers, using the i.i.d. assumption and the fact that $\mathbb{E}[X_i^2] \in (0, \infty)$ by Assumption 4. Similarly, $\|\mathbf{y}\|_2^2 \xrightarrow{P} \mathbb{E}[Y] \in [0, \infty)$.

Turning to α, we first see that by the Law of Large Numbers,

$$\frac{n(z)}{n} \xrightarrow{P} \mathbb{P}(Z = z)$$

for $n \to \infty$ and $z \in E$. Then observe that for any $z \in E$

$$|\hat{n}(z) - n(z)| \leqslant \sum_{k=1}^{n} |1(\hat{z}_k = z) - 1(z_k = z)| \leqslant \sum_{k=1}^{n} 1(\hat{z}_k \neq z_k) \leqslant n\delta.$$

Since $\delta \xrightarrow{P} 0$, also

$$\frac{\hat{n}(z)}{n} \xrightarrow{P} \mathbb{P}(Z = z),$$

thus

$$\alpha = \frac{n_{\min}}{n} = \min\left\{\frac{n(1)}{n}, \ldots, \frac{n(K)}{n}, \frac{\hat{n}(1)}{n}, \ldots, \frac{\hat{n}(K)}{n}\right\} \xrightarrow{P} \min_{z \in E} \mathbb{P}(Z = z) \in (0, \infty).$$

We finally consider ρ, and to this end we first see that

$$\frac{1}{n}\|(I - P_{\mathbf{Z}})\mathbf{x}_i\|_2^2 = \frac{1}{n}\sum_{k=1}^{n}(x_{i,k} - \overline{\mu}(z_k))^2 \xrightarrow{P} \mathbb{E}\left[\sigma_i^2(Z)\right] \in (0, \infty).$$

Moreover, using Lemma 1,

$$\begin{aligned}\left|\|(I - P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2 - \|(I - P_{\mathbf{Z}})\mathbf{x}_i\|_2^2\right| &= \left|\|(P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}})\mathbf{x}_i\|_2^2 + 2|\langle (I - P_{\hat{\mathbf{Z}}})\mathbf{x}_i, (P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}})\mathbf{x}_i\rangle\right| \\ &\leqslant \|P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}}\|_2^2\|\mathbf{x}_i\|_2^2 + 2\|P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}}\|_2\|\mathbf{x}_i\|_2^2 \\ &\leqslant \left(\frac{2\delta}{\alpha} + \sqrt{\frac{2\delta}{\alpha}}\right)\|\mathbf{x}_i\|_2^2.\end{aligned}$$

Hence

$$\rho \xrightarrow{P} \frac{\mathbb{E}\left[\sigma_i^2(Z)\right]}{\mathbb{E}[X_i^2]} \in (0, \infty).$$

Combining the limit results,

$$L_n \xrightarrow{P} L = \frac{2\sqrt{2}\mathbb{E}[X_i^2]^2}{\mathbb{E}\left[\sigma_i^2(Z)\right]^2\sqrt{\min_{z \in E}\mathbb{P}(Z = z)}}\sqrt{\frac{\mathbb{E}[Y^2]}{\mathbb{E}[X_i^2]}} \in (0, \infty).$$

To complete the proof, suppose first that $\frac{\operatorname{sep}(p)}{n} \to \infty$. Then

$$\sqrt{n}|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \leqslant \sqrt{n\delta}L_n$$

By (33) we have, under the assumptions given in case (2) of the theorem, that $n\delta \xrightarrow{P} 0$, and case (2) follows.

Finally, in the sub-Gaussian case, and if just $h_n = \frac{\operatorname{sep}(p)}{\log(n)} \to \infty$, then we can replace (33) by the bound

$$\mathbb{E}[\delta] \leqslant K\exp\left(-\frac{\operatorname{sep}(p)}{50v_{\max}}\right) + K^2\max_{z \neq v}\mathbb{P}\left(R_{z,v}^{(p)} > \tfrac{1}{10}\right).$$

Multiplying by n, we get that the first term in the bound equals

$$\begin{aligned}Kn\exp\left(-\frac{\operatorname{sep}(p)}{50v_{\max}}\right) &= K\exp\left(-\frac{\operatorname{sep}(p)}{50v_{\max}} + \log(n)\right) \\ &= K\exp\left(\log(n)\left(1 - \frac{h_n}{50v_{\max}}\right)\right) \to 0\end{aligned}$$

for $n \to \infty$. We conclude that the relaxed growth condition on p in terms of n in the sub-Gaussian case is enough to imply $n\delta \xrightarrow{P} 0$, and case (3) follows.

By the decomposition

$$\sqrt{n}(\widehat{\beta}_i^{\text{sub}} - \beta_i) = \sqrt{n}(\widehat{\beta}_i^{\text{sub}} - \widehat{\beta}_i) + \sqrt{n}(\hat{\beta}_i - \beta_i)$$

it follows from Slutsky's theorem that in case (2) as well as case (3),

$$\sqrt{n}(\widehat{\beta}_i^{\text{sub}} - \beta_i) = \sqrt{n}(\widehat{\beta}_i - \beta_i) + o_P(1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, w_i^2).$$

□

Appendix B. Gaussian mixture models

This appendix contains an analysis of a latent variable model with a finite E, similar to the one given by Assumption 4, but with Assumption 4(1) strengthened to

$$X_i \mid Z = z \sim \mathcal{N}(\mu_i(z), \sigma_i^2(z)).$$

Assumptions 4(2), 4(3) and 4(4) are dropped, and the purpose is to understand precisely when Assumption 2(2) holds in this model. That is, when Z can be recovered from $\mathbf{X}_{-i}$. To keep notation simple, we will show when Z can be recovered from $\mathbf{X}$, but the analysis and conclusion is the same if we left out a single coordinate.

The key to this analysis is a classical result due to Kakutani. As in Section 2, the conditional distribution of $\mathbf{X}$ given $Z = z$ is denoted P_z, and the model assumption is that

$$P_z = \bigotimes_{i=1}^{\infty} P_z^i \tag{34}$$

where P_z^i is the conditional distribution of X_i given $Z = z$. For Kakutani's theorem below we do not need the Gaussian assumption; only that P_z^i and P_v^i are equivalent (absolutely continuous w.r.t. each other), and we let $\frac{\mathrm{d}P_z^i}{\mathrm{d}P_v^i}$ denote the Radon-Nikodym derivative of P_z^i w.r.t. P_v^i.

Theorem 3 (Kakutani (1948))**.** *Let $z, v \in E$ and $v \neq z$. Then P_z and P_v are singular if and only if*

$$\sum_{i=1}^{\infty} -\log \int \sqrt{\frac{\mathrm{d}P_z^i}{\mathrm{d}P_v^i}}\, \mathrm{d}P_v^i = \infty. \tag{35}$$

Note that

$$\mathrm{BC}_{z,v}^i = \int \sqrt{\frac{\mathrm{d}P_z^i}{\mathrm{d}P_v^i}}\, \mathrm{d}P_v^i$$

is known as the Bhattacharyya coefficient, while $-\log\left(\mathrm{BC}_{z,v}^i\right)$ and $\sqrt{1 - \mathrm{BC}_{z,v}^i}$ are known as the Bhattacharyya distance and the Hellinger distance, respectively, between P_z^i and P_v^i. Note also that if $P_z^i = h_z^i \cdot \lambda$ and $P_v^i = h_v^i \cdot \lambda$ for a reference measure λ, then

$$\mathrm{BC}_{z,v}^i = \int \sqrt{h_z^i h_v^i}\, \mathrm{d}\lambda.$$

Proposition 5. *Let P_z^i be the $\mathcal{N}(\mu_i(z), \sigma_i^2(z))$-distribution for all $i \in \mathbb{N}$ and $z \in E$. Then P_z and P_v are singular if and only if either*

$$\sum_{i=1}^{\infty} \frac{(\mu_i(z) - \mu_i(v))^2}{\sigma_i^2(z) + \sigma_i^2(v)} = \infty \qquad \text{or} \tag{36}$$

$$\sum_{i=1}^{\infty} \log\left(\frac{\sigma_i^2(z) + \sigma_i^2(v)}{2\sigma_i(z)\sigma_i(v)}\right) = \infty \tag{37}$$

Proof. Letting $\mu = \mu_i(z)$, $\nu = \mu_i(v)$, $\tau = 1/\sigma_i(z)$ and $\kappa = 1/\sigma_i(v)$ we find

$$\begin{aligned}
\mathrm{BC}_{z,v}^i &= \int \sqrt{\frac{\tau}{\sqrt{2\pi}} \exp\left(-\frac{\tau^2}{2}(x-\mu)^2\right) \frac{\kappa}{\sqrt{2\pi}} \exp\left(-\frac{\kappa^2}{2}(x-\nu)^2\right)}\,\mathrm{d}x \\
&= \sqrt{\frac{\tau\kappa}{2\pi}} \int \exp\left(-\frac{(\tau^2+\kappa^2)x^2 - 2(\tau^2\mu + \kappa^2\nu)x + (\tau^2\mu^2 + \kappa^2\nu^2)}{4}\right)\mathrm{d}x \\
&= \sqrt{\frac{\tau\kappa}{2\pi}} \sqrt{\frac{4\pi}{\tau^2+\kappa^2}} \exp\left(\frac{(\tau^2\mu + \kappa^2\nu)^2}{4(\tau^2+\kappa^2)} - \frac{\tau^2\mu^2 + \kappa^2\nu^2}{4}\right) \\
&= \sqrt{\frac{2\tau\kappa}{\tau^2+\kappa^2}} \exp\left(-\frac{\tau^2\kappa^2(\mu-\nu)^2}{4(\tau^2+\kappa^2)}\right) \\
&= \sqrt{\frac{2\sigma_i(z)\sigma_i(v)}{\sigma_i^2(z) + \sigma_i^2(z)}} \exp\left(-\frac{(\mu_i(z) - \mu_i(v))^2}{4(\sigma_i^2(z) + \sigma_i^2(z))}\right).
\end{aligned}$$

Thus

$$\sum_{i=1}^{\infty} -\log\left(\mathrm{BC}_{z,v}^i\right) = \frac{1}{2}\sum_{i=1}^{\infty} \log\left(\frac{\sigma_i^2(z) + \sigma_i^2(v)}{2\sigma_i(z)\sigma_i(v)}\right) + \frac{1}{4}\sum_{i=1}^{\infty} \frac{(\mu_i(z) - \mu_i(v))^2}{\sigma_i^2(z) + \sigma_i^2(v)},$$

and the result follows from Theorem 3. □

Corollary 1. *Let P_z^i be the $\mathcal{N}(\mu_i(z), \sigma_i^2(z))$-distribution for all $i \in \mathbb{N}$ and $z \in E$. There is a mapping $f : \mathbb{R}^{\mathbb{N}} \to E$ such that $Z = f(\mathbf{X})$ almost surely if and only if either* (36) *or* (37) *holds.*

Proof. If either (36) or (37) holds, P_z and P_v are singular whenever $v \neq z$. This implies that there are measurable subsets $A_z \subseteq \mathbb{R}^{\mathbb{N}}$ for $z \in E$ such that $P_z(A_z) = 1$ and $P_v(A_z) = 0$ for $v \neq z$. Setting $A = \cup_z A_z$ we see that

$$P(A) = \sum_z P_z(A)\mathbb{P}(Z = z) = \sum_z P_z(A_z)\mathbb{P}(Z = z) = 1.$$

Defining the map $f : \mathbb{R}^{\mathbb{N}} \to E$ by $f(\mathbf{x}) = z$ if $\mathbf{x} \in A_z$ (and arbitrarily on the complement of A) we see that $f(\mathbf{X}) = Z$ almost surely.

On the other hand, if there is such a mapping f, define $A_z = f^{-1}(\{z\})$ for all $z \in E$. Then $A_z \cap A_v = \varnothing$ for $v \neq z$ and

$$\begin{aligned}
P_z(A_z) &= \frac{\mathbb{P}(\mathbf{X} \in A_z, Z = z)}{\mathbb{P}(Z = z)} = \frac{\mathbb{P}(f(\mathbf{X}) = z, Z = z)}{\mathbb{P}(Z = z)} \\
&= \frac{\mathbb{P}(f(\mathbf{X}) = Z, Z = z)}{\mathbb{P}(Z = z)} = \frac{\mathbb{P}(Z = z)}{\mathbb{P}(Z = z)} = 1.
\end{aligned}$$

Similarly, for $v \neq z$

$$P_v(A_z) = \frac{\mathbb{P}(\mathbf{X} \in A_z, Z = v)}{\mathbb{P}(Z = v)} = \frac{\mathbb{P}(f(\mathbf{X}) = z, Z = v)}{\mathbb{P}(Z = v)}$$
$$= \frac{\mathbb{P}(f(\mathbf{X}) \neq Z, Z = v)}{\mathbb{P}(Z = v)} = \frac{0}{\mathbb{P}(Z = v)} = 0.$$

This shows that P_z and P_v are singular, and by Proposition 5, either (36) or (37) holds. □

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. & Telgarsky, M. (2014), 'Tensor decompositions for learning latent variable models', *Journal of Machine Learning Research* **15**, 2773–2832.

Berk, R., Buja, A., Brown, L., George, E., Kuchibhotla, A. K., Su, W. & Zhao, L. (2021), 'Assumption lean regression', *The American Statistician* **75**(1), 76–84.

Ćevid, D., Bühlmann, P. & Meinshausen, N. (2020), 'Spectral deconfounding via perturbed sparse linear models', *Journal of Machine Learning Research* **21**(232), 1–41.

Cochran, W. G. (1968), 'Errors of measurement in statistics', *Technometrics* **10**(4), 637–666.

Durbin, J. (1954), 'Errors in variables', *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* **22**(1/3), 23–32.

D'Amour, A. (2019), On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', PMLR, pp. 3478–3486.

Gandhi, K. & Borns-Weil, Y. (2016), 'Moment-based learning of mixture distributions'.

Grimmer, J., Knox, D. & Stewart, B. (2023), 'Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding', *Journal of Machine Learning Research* **24**(182), 1–70.

Guo, B., Nie, J. & Yang, Z. (2022), 'Learning diagonal Gaussian mixture models and incomplete tensor decompositions', *Vietnam Journal of Mathematics* **50**, 421–446.

Guo, Z., Ćevid, D. & Bühlmann, P. (2022), 'Doubly debiased lasso: High-dimensional inference under hidden confounding', *The Annals of Statistics* **50**(3), 1320 – 1347.

Heinrich, P. & Kahn, J. (2018), 'Strong identifiability and optimal minimax rates for finite mixture estimation', *The Annals of Statistics* **46**(6A), 2844 – 2870.

Kakutani, S. (1948), 'On equivalence of infinite product measures', *Annals of Mathematics* **49**(1), 214–224.

Kalai, A. T., Moitra, A. & Valiant, G. (2012), 'Disentangling Gaussians', *Communications of the ACM* **55**(2), 113–120.

Kallenberg, O. (2021), *Foundations of modern probability*, Probability and its Applications (New York), third edn, Springer-Verlag, New York.

Leek, J. T. & Storey, J. D. (2007), 'Capturing heterogeneity in gene expression studies by surrogate variable analysis', *PLOS Genetics* **3**(9), 1–12.

Lundborg, A. R., Kim, I., Shah, R. D. & Samworth, R. J. (2023), 'The projected covariance measure for assumption-lean variable significance testing', *arXiv:2211.02039* .

Miao, W., Hu, W., Ogburn, E. L. & Zhou, X.-H. (2023), 'Identifying effects of multiple treatments in the presence of unmeasured confounding', *Journal of the American Statistical Association* **118**(543), 1953–1967.

Ndaoud, M. (2022), 'Sharp optimal recovery in the two component Gaussian mixture model', *The Annals of Statistics* **50**(4), 2096 – 2126.

Ogburn, E. L., Shpitser, I. & Tchetgen, E. J. T. (2020), 'Counterexamples to "the blessings of multiple causes" by Wang and Blei', *arXiv:2001.06555* .

Patterson, N., Price, A. L. & Reich, D. (2006), 'Population structure and eigenanalysis', *PLOS Genetics* **2**(12), 1–20.

Pearl, J. (2009), *Causality*, Cambridge University Press.

Peters, J., Janzing, D. & Schölkopf, B. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, Cambridge, MA, USA.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics* **38**(8), 904–909.

Schennach, S. M. (2016), 'Recent advances in the measurement error literature', *Annual Review of Economics* **8**, 341–377.

Song, M., Hao, W. & Storey, J. D. (2015), 'Testing for genetic associations in arbitrarily structured populations', *Nat Genet* **47**(5), 550–554.

Stewart, G. W. (1977), 'On the perturbation of pseudo-inverses, projections and linear least squares problems', *SIAM Review* **19**(4), 634–662.

van der Vaart, A. W. (1998), *Asymptotic statistics*, Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.

Vansteelandt, S. & Dukes, O. (2022), 'Assumption-lean inference for generalised linear model parameters', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84**(3), 657–685.

Wang, J., Zhao, Q., Hastie, T. & Owen, A. B. (2017), 'Confounder adjustment in multiple hypothesis testing', *Ann. Statist.* **45**(5), 1863–1894.

Wang, Y. & Blei, D. M. (2019), 'The blessings of multiple causes', *Journal of the American Statistical Association* **114**(528), 1574–1596.

Wang, Y. & Blei, D. M. (2020), 'Towards clarifying the theory of the deconfounder', *arXiv:2003.04948* .

3.1 Additional Discussion

In [Adjustment], our analysis is largely agnostic to causal interpretation. In this section, we will discuss a causal interpretation of the model in greater detail. First, we introduce and motivate the main assumptions of Wang and Blei [2019]. Next, we discuss two main criticisms of the original paper. The first is the original claim that $\sigma(Z)$ did not need to be unique for deconfounding to be possible. The second is the debate whether a deconfounder can pick up on a mediator.

3.1.1 The Assumptions

As presented in Wang and Blei [2019], the deconfounder originally rested on three assumptions, which we discuss here.

As pointed out by Grimmer et al. [2023], substitute adjustment essentially coincides with a semiparametric partially linear regression model

$$\mathbb{E}[Y|X_1 = x_1, ..., X_p = x_p] = \beta_0 + \beta_1 x_1 + h(x_2, ..., x_p).$$

However, we disagree that this observation makes the deconfounder idea unnecessary. For such a partially linear regression to be feasible, we desire a principled way of restricting the function class of h. Assumption 1 will give us such a principled restriction.

Assumption 1 (Conditional Independence)**.** *The entries of X are mutually independent given Z.*

This is exactly [Adjustment] Assumption 2(1). Interpreted causally, Assumption 1 says that the treatments are causally non-adjacent, and that the dependency structure of X is entirely explained by a hidden common cause, Z. However, understood as a restriction on the function class of h, it says that h must be of the form $g \circ f$, where $f(X_2, ..., X_p)$ is required to make $X_2, ..., X_p$ independent. In cases where the causal model is plausible, this could be a principled choice of function class.

However, Assumption 1 is not really a restriction on Z yet—X is mutually independent conditional on X, but we certainly don't want to adjust X for X. Intuitively, the idea of Wang and Blei [2019] is to adjust for as little information as necessary to make X independent, but no more. Essentially, the only information the causal model gives us is that X are conditionally independent given the true Z. Therefore, it is sensible to reconstruct Z using all information in the *dependency* structure of X; however, we are not justified in using any information from the marginal distributions, because we cannot know what part is idiosyncratic to X_i and what part is due to a confounder. Therefore, in the original paper, Wang and Blei [2019] propose

Assumption 2 (Minimality)**.** *For any random variable Z', if Z' satisfies Assumption 1, then $\sigma(Z) \not\supset \sigma(Z')$.*

The remarkable thing about Assumption 2 is that, if it were sufficient to control for confounding, then it would not be necessary to recover the σ-algebra of the true confounder Z; it would be sufficient to adjust for *any* minimal σ-algebra which makes X

conditionally independent. However, this assumption is too weak to control for unobserved confounding in general; problems can occur when the σ-algebra satisfying Assumption 2 is not unique, as we will discuss this in Section 3.1.2. Wang and Blei [2020] resolve this problem by assuming there is a unique map $f : \mathbb{R}^p \to E$ such that $f(X)$ satisfies Assumptions 1 and 2.

In contrast, any random variable satisfying [Adjustment] Assumption 2(2) is sufficient to adjust for confounding due to [Adjustment] Proposition 1. That assumption is in the same spirit as Assumption 2: informally, Assumption 2(2) entails that any information in Z must be in at least two X_i, so that it is actually required to render X independent.

For substitute adjustment to actually remove confounding bias, the recovered variable must actually relate to the confounding. Since the substitute is constructed only from the dependency structure of X and not from the marginal distributions, a confounder which affects only one treatment variable cannot be captured by the substitute. This motivates Wang and Blei [2019] to rule out single-cause confounders:

Assumption 3 (Single Ignorability)**.** *For all $i \in \{1, ..., d\}$, $x \in \mathcal{X}$, the potential outcome $Y^{X_i=x} \perp\!\!\!\perp X_i | Z$.*

This assumption is weaker than the usual assumption of no unobserved confounding.

3.1.2 Uniqueness of Z

Wang and Blei [2019] claimed that *any* substitute satisfying Assumptions 1 to 3 could be used for adjustment. This promised to allow for unbiased treatment effect estimation under weaker assumptions than standard unconfoundedness, and to allow model checking of the substitute confounder.

The problem is that without some way to chose between different minimal $\sigma(Z)$, the linear regression parameter β (or any other functional of the counterfactual distribution) is not generally identifiable. As Ogburn et al. [2020] point out, a variable Z satisfying Assumption 1 controls for confounding only insofar as the dependency structure on X relates to the confounding between X and $Y^{X=x}$—i.e. insofar as the *same* variable Z satisfies both Assumption 1 and Assumption 3. D'Amour [2019] demonstrate this with a specific theoretical example. Similar to our analysis for Gaussian mixtures in [Adjustment]Appendix B, their example further reminds us that $p \to \infty$ is not sufficient to recover Z, because this in itself does not necessarily entail mutually singular conditional distributions $X|Z$. It is in light of these critiques that Wang and Blei [2020] strengthened Assumption 2.

But even with this correction, we must remember that these assumptions are just assumptions: D'Amour [2019] provide a simple copula argument showing that confounding is not generally discoverable from X alone. Only when causal relations between treatments are negligible (Assumption 1) and when the confounding structure between X and Y is related to the dependency structure within X (Assumption 3) does substitute adjustment actually adjust for confounding.

3.1.3 Picking up a Mediator, Picking up a Mechanism

When the assumptions of Wang and Blei [2020] hold in finite dimensions, the true confounder Z is not only causally prior to X, but also can be viewed as a deterministic function of the treatments X. This could raise many conceptual worries. For example, since Z is X-measurable, it follows that $Y^{X_i=x} \perp\!\!\!\perp Z|X$, indicating that Z is not a confounder according to usual definitions of confounding; "Confounders confound *because* they are related to potential outcomes even conditional on the observed treatment and outcome." [Ogburn et al., 2019] It is, in part, in response to this concern that we consider an infinite dimensional X in [Adjustment]; we only ever observe finitely many treatments, so that it still makes sense to talk about Z as a latent confounder.

The pinpointedness of Z also indicates a type of multicolinearity problem in the full deconfounder regression (in which the recovered substitute is merely appended to the full design matrix). This problem was investigated by [Grimmer et al., 2023]. For example, if $\hat{Z}$ is constructed from X by linear factor analysis, then it is colinear with X; thus, an ordinary least squares regression of Y on X and $\hat{Z}$ is undefined. More generally, consider a non-parametric regression

$$\mathbb{E}[Y|X_1 = x_1, ..., X_p = x_p] = b(X) + g(Z) \tag{1}$$

where b and g are restricted to some function classes. We must be sure, for example, the function classes are appropriately restricted so that b cannot first reconstruct Z as an intermediate step. [Ogburn et al., 2019] Clearly, we avoid this concern by considering each treatment individually in [Adjustment] Algorithms 1-3.

The multicolinearity problem is subtly related to the issue of "picking up a mediator." Lemma 4 of Wang and Blei [2019] claims that the substitute confounder cannot pick up a mediator— that is, no mediator is measurable with respect to the substitute confounder. If true, this is reassuring; it is unwise to adjust for mediators. However, it seems that the lemma is true, but surprisingly uninteresting. More interesting is the question, "can the substitute confounder pick up a mechanism"? The answer seems to be yes.

First a discussion of Lemma 4. The proof reveals that the claim relies on a specific but widespread understanding of the term "mediator". For example, Imai et al. [2010] requires that a mediator M satisfy $M \not\perp\!\!\!\perp Y^{X=x}|X$ for some potential treatment level x. The implication is that, by definition, mediators exclude X-measurable variables; and clearly every $\hat{Z}$-variable variable is also X-measurable because $\hat{Z} = f(X)$. In this sense, Lemma 4 of Wang and Blei [2019] is correct, but trivial. (X-measurability of the substitute confounder is indeed implicit in Lemma 4; otherwise the proof given in Wang and Blei [2019] fails, since it is possible that $Z = Y^{X=x}$ satisfies $X \perp\!\!\!\perp Z$ but $Z \not\perp\!\!\!\perp Y^{X=x}|X$, which would contradict the proof.)

A similar argument addresses concerns about M-bias and single-cause colliders raised by Ogburn et al. [2019]. However, this does not automatically greenlight adjustment with respect to $\hat{Z}$, precisely because the mechanism by which X causes Y could be similar to the true function f which pinpoints Z from X. Thus the possible issue is not that we are adjusting for a mediator, but that the true mechanisms align.

This is actually a major concern of Ćevid et al. [2020]: in linear models, the PCA correction often forces $\hat{\beta}$ orthogonal to the largest principle components, which is catastrophic when the true treatment mechanism is approximately colinear with the true linear factor loadings. Perhaps $\hat{Z}$ cannot pick up a mediator, but it can certainly pick up a mechanism. Ćevid et al. [2020] avoid their version of the problem by assuming sparse β and dense confounding, which forces orthogonality; Wang and Blei [2019] avoid picking up a mechanism by the function-class restrictions in Theorem 6, and the overlap assumption of Theorem 7. We avoid this and related problems by performing assumption lean substitute adjustment on each dimension of X individually.

4 Causal Interpretations of Lévy-driven Ornstein Uhlenbeck Processes

This chapter contains the following paper:

[Precision] [Recke et al., 2024]. C. O. Recke, J. Adams, and N. R. Hansen. Non-Gaussian graphical precision models. 2024.

In [Precision], we show that under a condition we call drift-volatility balance causal conclusions can easily be drawn from the steady-state observational distribution of Ornstein Uhlenbeck processes. Furthermore, we derive equations relating higher-order cumulants of the steady-state observational distribution to the drift- and volatility-parameters of the Ornstein-Uhlenbeck process. From these equations, we show that drift-volatility balance is a falsifiable property.

The proof of [Precision] Proposition 3.1 is very concise, but also very abstract. In Section 4.1, we provide a second proof of this proposition in the special case where the dependence structure of the Lévy process Z is explained by a linear mixing of arbitrarily many independent Lévy processes. Because this second proof is from first principles about the time dynamics of the stochastic process, it provides a mechanistic intuition that is absent in the more concise proof. Moreover, we show that under this additional linearity assumption, the cumulants of Z admit a symmetric tensor decomposition.

NON-GAUSSIAN GRAPHICAL PRECISION MODELS

CECILIE OLESEN RECKE, JEFFREY ADAMS, AND NIELS RICHARD HANSEN

This draft manuscript represents work in progress.

ABSTRACT. Sparse estimation of precision matrices is widely used also beyond the Gaussian case. Since a vanishing partial correlation does not imply a conditional independence for non-Gaussian data, it may, however, be difficult to give a proper interpretation of the sparsity pattern of such a precision matrix. We give a novel interpretation in terms of *operator selfdecomposable* (OSD) distributions, which appear as steady-state distributions for a class of Markov processes. A sparse precision matrix is then interpretable in terms of the process dynamics whenever the dynamics satisfies a condition we term *drift-volatility balance*. In the Gaussian case, the condition is equivalent to detailed balance, which results in a classical Gaussian graphical model. If drift-volatility balance is not satisfied, the precision matrix will generally be dense. We derive equations for the higher order cumulants of the non-Gaussian OSD distributions. These equations allow us to derive a rank constraint that holds under drift-volatility balance, and which is expressible in terms of the second and third order cumulants.

1. INTRODUCTION

We consider a multivariate random variable $X \in \mathbb{R}^p$ with finite second moment. We denote its covariance matrix by Σ, and when it is invertible we denote the corresponding precision matrix by $\Theta = \Sigma^{-1}$.

If $G = ([p], E)$ is an undirected graph with nodes $[p] = \{1, \ldots, p\}$ and edges E we let $\mathbb{R}^E$ denote the symmetric $p \times p$ matrices whose (i, j)-th entry is non-zero only if $\{i, j\} \in E$, and

$$\mathrm{PD}_G = \{\Theta \in \mathbb{R}^E \mid \Theta \text{ is positive definite}\}.$$

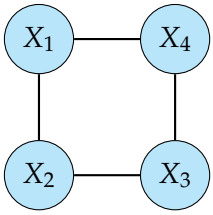
The set PD_G constitutes a graphical precision model given by the graph G, and any $\Theta \in \mathrm{PD}_G$ inherits the sparsity pattern of (the adjacency matrix of) G. If $X \sim \mathcal{N}(0, \Sigma)$ has a Gaussian distribution with $\Theta = \Sigma^{-1} \in \mathrm{PD}_G$ the distribution of X factorizes w.r.t. G, whence the distribution satisfies the global Markov property w.r.t. G, and separation in G implies a corresponding conditional independence among coordinates of X. In particular, for Gaussian distributions

$$\Theta_{ij} = 0 \quad \Longleftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid (X_k)_{k \neq i,j}, \tag{1}$$

see Proposition 5.2 in [Lau96].

We will consider the graphical precision model PD_G for non-Gaussian distributions, and we are particularly interested in (semiparametric) models that

(1) can explain the particular sparsity pattern of Θ as encoded by G
(2) and come with a testable causal interpretation



$$\Theta = \begin{pmatrix} * & * & 0 & * \\ * & * & * & 0 \\ 0 & * & * & * \\ * & 0 & * & * \end{pmatrix}$$

FIGURE 1. The graphical precision model PD_G for G the (undirected) four-cycle graph is equivalent to the constraints $\Theta_{13} = \Theta_{31} = 0$ and $\Theta_{24} = \Theta_{42} = 0$ for all $\Theta \in \mathrm{PD}_G$.

We will, nevertheless, first discuss the case where the distribution of X is Gaussian and where $\Theta \in \mathrm{PD}_G$ implies a range of factorization and conditional independency properties about the joint distribution of X. These properties do, however, not explain the origin of G. That is, they do not explain the mechanisms that can generate such a multivariate Gaussian distribution.

The graphical model with $p = 4$ given by the four-cycle, see Figure 1, is an interesting example; it is the simplest example of an undirected graphical independence model that cannot be represented by a DAG [Fry90, Theorem 5.6], whence the four-cycle cannot be (faithfully) explained by a set of recursive regressions. Cox and Wermuth [CW00] investigated other possible mechanisms that can explain the four-cycle, including limits and steady-state distributions for stochastic dynamical systems. The four-cycle – and any other Gaussian graphical model for that matter – can be explained as the steady-state distribution of a Gaussian Markov process, see also [LR02]. In fact, there is an entire family of such possible explanations, but among these it is only for the reversible Gaussian processes (those that satisfy detailed balance) that we can link the sparsity of Θ to the causal interpretation encoded by the dynamics of the process.

For non-Gaussian distributions we must be even more careful when interpreting the sparsity pattern of the precision matrix since (1) no longer holds. In general, $\Theta_{ij} = 0$ is only equivalent to a vanishing partial correlation between X_i and X_j, not a conditional independence, and even if $\Theta \in \mathrm{PD}_G$, we cannot use separation in G to infer conditional independencies. However, we argue that a sparse precision matrix maintains a natural interpretation for data from a steady-state distribution of a Markov process – whenever this process satisfies a condition we call *drift-volatility balance*. The steady-state distributions we consider are known as *operator selfdecomposable* distributions, with the Gaussian distribution being a special case. In the Gaussian case, drift-volatility balance is equivalent to detailed balance. Using the causal interpretation entailed by the Markov process we show that under the drift-volatility balance condition, interventional means and covariances are computable from Θ even in the non-Gaussian case, which is a generalization of Proposition 5 in [LR02].

In practice, given n i.i.d. observations, we are interested in estimating Θ. We are particularly interesting in interpreting the estimated Θ in terms of a steady-state distribution – and possibly draw causal inference. Estimation of a precision matrix is thoroughly investigated, and sparsity of Θ is a common assumption to ensure efficient estimation of either $\Theta \in \mathrm{PD}_G$ or of G itself, notably in the high-dimensional case where $p \gg n$. The graphical lasso [YL07; BGd08; FHT08] based on the Gaussian log-likelihood is one popular example of an estimator that enforces a sparsity constraint on Θ – in this case via

a 1-norm penalty. Conditions that ensure high-dimensional consistency of the graphical lasso even in the non-Gaussian case are well known [Rav+11]. Alternative estimators of Θ based on multivariate t-distributions [FD11] and elliptical distributions [VF11] have been considered as well, see also [DM17] for a detailed review of graph estimation for Gaussian as well as non-Gaussian data. There are thus multiple well studied estimators of sparse precision matrices also for non-Gaussian data.

Whether data is Gaussian or non-Gaussian, it remains important to investigate if a sparse precision matrix is an appropriate model. When Θ is the precision of a steady-state distribution we argue that sparsity of Θ is only really explainable by the dynamics of the process under drift-volatility balance. A sparse estimate of Θ is, however, no evidence in itself of drift-volatility balance, and a causal interpretation without further considerations is dubious. In the Gaussian case, drift-volatility balance is untestable without additional structural assumptions, but in the non-Gaussian case, drift-volatility balance has certain implications for the higher order moments. We derive equations characterizing all higher order cumulants of the operator selfdecomposable distributions. We use this to construct a matrix from Θ and the third order cumulant tensor, which is rank deficient under drift-volatility balance. Its empirical version can then be used to test the hypothesis of drift-volatility balance.

2. Operator selfdecomposable distributions

2.1. **OSD distributions and Lévy processes.** Recall that a continuous time stochastic process $Z = (Z_t)_{t\geq 0}$ with $Z_t \in \mathbb{R}^p$ is a Lévy process if: $Z_0 = 0$; if the increments of Z are independent and stationary; and if Z is continuous in probability. Letting $(\Delta_\delta Z)_s = Z_{s\delta} - Z_{(s-1)\delta}$ for a $\delta > 0$ it holds that $(\Delta_\delta Z)_1, \ldots, (\Delta_\delta Z)_n$ are i.i.d. and $Z_{n\delta} = \sum_{s=1}^{n} (\Delta_\delta Z)_s$ is a random walk. Fixing $\delta > 0$ and $\rho \in (0,1)$, the weighted sum

$$X = \sum_{s=0}^{\infty} \rho^s (\Delta_\delta Z)_s$$

converges almost surely. The distribution of X on $\mathbb{R}^p$ is the invariant distribution of the discrete time Markov process given by the autoregression

$$X_{t+1} = \rho X_t + (\Delta_\delta Z)_t.$$

The operator selfdecomposable distributions are defined below by a representation analogous to the infinite weighted sum above. In this definition, the sum is replaced by an integral corresponding to taking the limit $\delta \to 0$, and the weight $\rho^s = e^{s\log(\rho)}$ is replaced by the linear operator e^{sM} for a $p \times p$ matrix M, which entails that the different coordinates of Z are also mixed together. The condition that $\log(\rho) < 0$ is replaced by the condition that all eigenvalues of M have strictly negative real part. Such a matrix is called a *stable matrix*. To ensure convergence of the integral (3) below, we need to assume that

$$E(\log(1 + \|Z_1\|)) < \infty. \tag{2}$$

We will throughout assume that all Lévy processes considered fulfill this unrestrictive integrability condition.

Definition 2.1 (OSD distributions)**.** Let M be a $p \times p$ stable matrix and let $Z = (Z_t)_{t\geq 0}$ denote a p-dimensional Lévy process satisfying (2). The distribution of

$$X = \int_0^\infty e^{sM} \mathrm{d}Z_s \tag{3}$$

is called M-selfdecomposable. A distribution is operator selfdecomposable, or OSD, if it is M-selfdecomposable for some stable M and some Lévy process Z.

Operator selfdecomposable distributions where first studied by Urbanik [Urb72] under the name "Lévy's probability measures". Traditionally, X is defined to have an M-selfdecomposable distribution if there for all $t > 0$ exists X_t independent of X such that

$$X \overset{\mathcal{D}}{=} e^{tM} X + X_t. \tag{4}$$

It is not so difficult to show that an X given by (3) satisfies (4). Independently, Jurek [Jur82] and Wolfe [Wol82] showed the other direction, thus M-selfdecomposability is equivalently defined by the distributional property (4) and by the representation (3) for some Lévy process Z. See also [SY84] for several alternative characterizations of OSD distributions. We take (3) as our definition since this representation is directly useful for the results we show.

It is worth explaining the term "operator selfdecomposable". Historically, if $X \overset{\mathcal{D}}{=} X' + X''$ for independent X' and X'' the distribution of X is said to be *decomposable*. If it is possible to take $X' = \rho X$ in this decomposition for any $\rho \in (0, 1)$, the distribution of X is said to be *selfdecomposable*. The M-selfdecomposable distributions are by (4) a generalization where $\rho \in (0, 1)$ is replaced by the linear operator e^{tM} for $t > 0$. The selfdecomposable distributions are thus precisely those that are I-selfdecomposable with I the $p \times p$ identity matrix. Examples of OSD distributions are the multivariate stable distributions [SY84, Example 4.1], which include the Gaussian distributions. The larger class of multivariate generalized hyperbolic distributions [Mas04, Section 5], which include the multivariate t-distributions, are also OSD. On the other hand, the *infinite divisibility* of the Lévy process Z in (3) implies that any OSD is infinitely divisible, see Figure 2.

2.2. **A causal interpretation.** One main motivation for studying operator selfdecomposable distributions is the following result.

Proposition 2.2 ([SY84, Theorem 4.1])**.** *Let M be a $p \times p$ stable matrix and let $Z = (Z_t)_{t\geq 0}$ denote a p-dimensional Lévy process satisfying* (2)*. The M-selfdecomposable distribution given by* (3) *is the unique steady-state distribution of the stationary Markov process solving the SDE*

$$\mathrm{d}X_t = MX_t \mathrm{d}t + \mathrm{d}Z_t. \tag{5}$$

The fact that any OSD distribution can occur as the steady-state distribution of a Markov process solving (5) makes this class of distributions natural when considering cross-sectional data from a dynamical system. Moreover, the SDE will also allow us to define interventions and thus give a causal interpretation of OSD distributions.

Following [LR02] and [SH14], the SDE can be given a causal interpretation[1] with interventions defined by substitution. That is, we define the intervention on a coordinate by

[1]This is a *structural* interpretation; the structure of the SDE is assumed invariant to interventions, and the SDE is interpreted as an infinitesimal structural causal model, see [SH14].

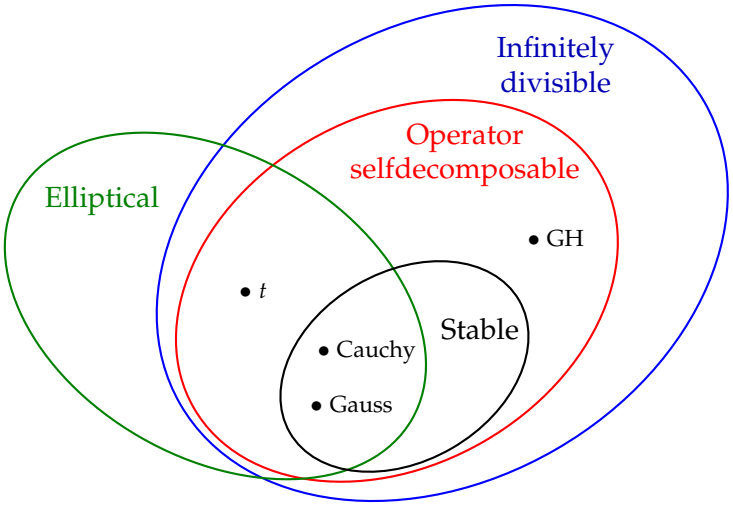


FIGURE 2. All stable distributions are OSD, and OSD distributions are infinitely divisible. There is some overlap between OSD and elliptical distributions. The generalized hyperbolic (GH) distributions are OSD, the t-distributions are, in addition, elliptical and the Cauchy as well as the Gaussian distributions are both stable and elliptical.

substituting that coordinate with a fixed value in the SDE. Interventions may, of course, be defined for any subset of coordinates similarly, and we can also define more general interventions where coordinates are not just fixed, see [SH14].

Here we consider intervening on a block of coordinates, and we let

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

denote a block partition of M and $X_t = (X_t^{(1)}, X_t^{(2)})$ and $Z_t = (Z_t^{(1)}, Z_t^{(2)})$ denote the corresponding partition of X_t and Z_t, respectively. Intervening, as defined in [LR02; SH14], by fixing $X_t^{(1)} = x^{(1)}$ for all $t \geq 0$ gives the SDE

$$\begin{aligned} \mathrm{d}X_t^{(2)} &= (M_{22}X_t^{(2)} + M_{21}x^{(1)})\mathrm{d}t + \mathrm{d}Z_t^{(2)} \\ &= M_{22}X_t^{(2)}\mathrm{d}t + \mathrm{d}(Z_t^{(2)} + M_{21}x^{(1)}t). \end{aligned} \tag{6}$$

Proposition 2.3. *If M_{22} is a stable matrix, the steady-state interventional distribution of $X_t^{(2)}$ is M_{22}-selfdecomposable. If $Z_1^{(2)}$, and hence $X^{(2)}$, has finite second moment then*

$$\mathbb{E}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) = -(M_{22})^{-1}(a^{(2)} + M_{21}x^{(1)}) \tag{7}$$

$$\mathbb{V}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) = \Sigma^{(2)} = \int_0^\infty e^{sM_{22}}C_{22}e^{sM_{22}^T}\mathrm{d}s \tag{8}$$

where $a^{(2)} = \mathbb{E}(Z_1^{(2)})$ and $C_{22} = \mathbb{V}(Z_1^{(2)})$.

Proof. When M_{22} is stable, the steady-state distribution of the solution to the interventional SDE (6) gives that the interventional distribution of $X^{(2)}$ is represented by

$$X^{(2)} = \int_0^\infty e^{sM_{22}} \mathrm{d}(Z_s^{(2)} + M_{21}x^{(1)}s) \tag{9}$$
$$= \int_0^\infty e^{sM_{22}} \mathrm{d}Z_s^{(2)} + \int_0^\infty e^{sM_{22}} \mathrm{d}s M_{21}x^{(1)}$$
$$= \int_0^\infty e^{sM_{22}} \mathrm{d}Z_s^{(2)} - (M_{22})^{-1}M_{21}x^{(1)}, \tag{10}$$

where we have used that when M_{22} is stable,

$$\int_0^\infty e^{sM_{22}} \mathrm{d}s = (M_{22})^{-1}e^{sM_{22}}\Big|_0^\infty = -(M_{22})^{-1}.$$

We note that (9) directly shows that the distribution of $X^{(2)}$ is M_{22}-selfdecomposable. When $Z^{(2)}$ has finite first moment, the expectation of (10) gives

$$\mathbb{E}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) = \int_0^\infty e^{sM_{22}}a^{(2)}\mathrm{d}s - (M_{22})^{-1}M_{21}x^{(1)}$$
$$= -(M_{22})^{-1}(a^{(2)} + M_{21}x^{(1)}).$$

If $Z^{(2)}$ has finite second moment, taking the covariance of (10) gives the formula (8); the computation is a special case of the proof of Proposition 3.1 below. □

Note that if M_{22} is *not* stable, the intervened SDE (6) will generally not have a steady-state distribution[2] and the interventional distribution is undefined.

When M_{22} is stable, the formulas (7) and (8) are superficially similar to the formulas for the (observational) conditional mean and covariance in the multivariate Gaussian distribution. The interventional mean is, e.g., an affine function of $x^{(1)}$ and the interventional covariance is independent of $x^{(1)}$. There are, however, important differences, and we emphasize that:

- In the general non-Gaussian case, (7) and (8) only give the first two interventional moments, which do *not* characterize the entire interventional distribution.
- Even in the Gaussian case, the interventional mean and covariance only coincide with the observational conditional mean and covariance in special cases, see Proposition 5 in [LR02].
- There are non-Gaussian examples where the observational conditional mean and covariance always differ from the interventional formulas (7) and (8). Thus Proposition 5 in [LR02] does not generalize from the Gaussian case to all OSD distributions, but see Section 4.1.

3. Cumulants

3.1. **The Lyapunov tensor equation.** We derive the general Lyapunov tensor equation for M-selfdecomposable distributions. We recall the notion of an n-mode product between a tensor and a matrix, which is also sometimes called the Tucker product.

[2]It might have if the distribution of Z_1 is degenerate but not for non-degenerate distributions of Z_1.

The n-mode product of a tensor $K \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $A \in \mathbb{R}^{J \times I_n}$ denoted $K \times_n A$ is a $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$ tensor with elementwise entries

$$(K \times_n A)_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} K_{i_1 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} A_{j i_n}.$$

Proposition 3.1. *Let M be a $p \times p$ stable matrix and let $Z = (Z_t)_{t \geq 0}$ denote a p-dimensional Lévy process with finite k-th moment. Then the corresponding M-selfdecomposable distribution given by* (3) *has finite k-th moment, and the k-th order cumulant tensor $K = \mathrm{cum}_k(X)$ solves the equation*

$$K \times_1 M + \ldots + K \times_k M + \mathcal{C}_k = 0 \tag{11}$$

where $\mathcal{C}_k = \mathrm{cum}_k(Z_1)$ is the k-order cumulant tensor of Z_1.

Proof. Using multilinearity of the cumulant operator,

$$\begin{aligned} K = \mathrm{cum}_k(X) &= \int_0^\infty \cdots \int_0^\infty \mathrm{cum}_k(\mathrm{d}Z_{s_1}, \ldots, \mathrm{d}Z_{s_k}) \times_1 e^{s_1 M} \times_2 e^{s_2 M} \ldots \times_k e^{s_k M} \\ &= \int_0^\infty \mathcal{C}_k \times_1 e^{sM} \times_2 e^{sM} \ldots \times_k e^{sM} \mathrm{d}s \end{aligned}$$

where we have used that for a Lévy process, the k-th order cumulant measure is "diagonal" and equals

$$\mathrm{cum}_k(\mathrm{d}Z_{s_1}, \ldots, \mathrm{d}Z_{s_k}) = \mathcal{C}_k \delta_{s_1, \ldots, s_k} H^1(ds_1, \ldots, ds_k)$$

with H^1 the 1-dimensional Hausdorff measure. That K solves (11) follows from Theorem 3.4 in [XW21]. □

The special case $k = 1$ gives the equation $M\mathbb{E}(X) + \mathbb{E}(Z_1) = 0$ and for $k = 2$ the second order cumulant equation results in the well known Lyapunov equation for the covariance matrix $\Sigma = \mathbb{V}(X) = \mathrm{cum}_2(X)$ in terms of M and the covariance matrix $C = \mathcal{C}_2 = \mathbb{V}(Z_1)$.

Corollary 3.2. *Let M be a $p \times p$ stable matrix and let $Z = (Z_t)_{t \geq 0}$ denote a p-dimensional Lévy process with finite second moment. Let $a = \mathbb{E}(Z_1)$ and $C = \mathbb{V}(Z_1)$. Then the corresponding OSD has mean $\xi = -M^{-1}a$ and covariance matrix Σ solving the continuous Lyapunov equation*

$$M\Sigma + \Sigma M^T + C = 0. \tag{12}$$

Remark 3.3. Note that Corollary 3.2 implies that the covariance matrix $\Sigma^{(2)}$ given by the integral representation in (8), and appearing there as the interventional covariance matrix, solves the Lyapunov equation

$$M_{22}\Sigma^{(2)} + \Sigma^{(2)} M_{22}^T + C_{22} = 0.$$

4. Drift-volatility balance

In this section we assume throughout that Z has independent coordinates[3] and finite second moment. Thus

$$C = \mathbb{V}(Z_1) = \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & c_p \end{pmatrix}. \tag{13}$$

4.1. **The drift-volatility balance condition.** We introduce a concept we call *drift-volatility balance* for Markov processes solving (5), which expresses a form of local flow balance between any two coordinates i and j. It is a weaker condition than detailed balance, which requires the Markov process to be reversible.

Except for the Gaussian case, solutions of (5) are generally not reversible even if drift-volatility balance holds, but drift-volatility balance is sufficient to ensure some very important links between the dynamics of the process and its steady-state distribution.

Definition 4.1 (Drift-volatility balance)**.** Let M be a stable matrix and C a diagonal matrix with diagonal elements $c_1, \dots, c_p > 0$. We say that (M, C) satisfies drift-volatility balance (DVB) if

$$M_{ij}c_j = M_{ji}c_i. \tag{14}$$

A Markov process solving (5) is likewise said to satisfy drift-volatility balance if (M, C) does so.

The drift-volatility balance condition can also be formulated as the matrix identity:

$$MC = CM^T, \tag{15}$$

which simply states that MC is symmetric.

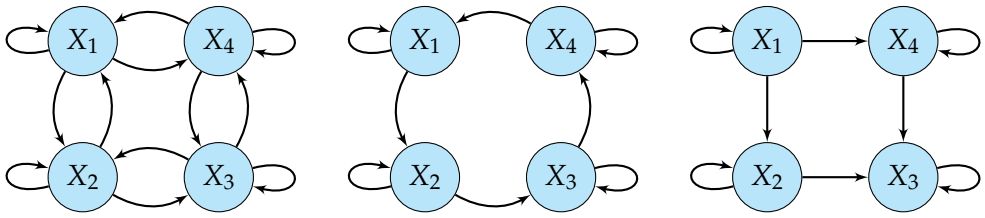
Proposition 4.2. *Let M be a stable matrix and C a positive definite diagonal matrix. If (M, C) satisfies drift-volatility balance then*

$$\Sigma = -\frac{1}{2}M^{-1}C \tag{16}$$

is the unique solution of the Lyapunov equation (12)*. Moreover, if $Z_t = C^{\frac{1}{2}}W_t$ for a standard Brownian motion W, the corresponding Gaussian Markov process solving* (5) *satisfies detailed balance with $\mathcal{N}(0, \Sigma)$ as steady-state distribution.*

Proof. Under the stated conditions, the Lyapunov equation has a unique solution, and when drift-volatility balance holds, (16) is directly seen to solve the Lyapunov equation. When Z_t is a Brownian motion the Markov process solving (5) is well known to be Gaussian with Gaussian transition densities and Gaussian steady-state distribution, and the detailed balance condition is easily verified from (16). □

[3]A slightly more general assumption, that might also work, is that $Z_t = DZ'_t$ for a $p \times q$ matrix D with orthogonal rows and Z' a q-dimensional Lévy process with independent coordinates.



$$M = \begin{pmatrix} * & * & 0 & * \\ * & * & * & 0 \\ 0 & * & * & * \\ * & 0 & * & * \end{pmatrix} \qquad M = \begin{pmatrix} * & 0 & 0 & * \\ * & * & 0 & 0 \\ 0 & * & * & 0 \\ 0 & 0 & * & * \end{pmatrix} \qquad M = \begin{pmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ 0 & * & * & * \\ * & 0 & 0 & * \end{pmatrix}$$

FIGURE 3. Three representations of the sparsity patterns of M-matrices using directed graphs. With $C = I$, M satisfies DVB if and only if M is symmetric, in which case $\Theta = \Sigma^{-1} = -2M$ (left). If M is not symmetric (middle and right), the sparsity pattern of $\Theta = \Sigma^{-1}$, with Σ solving the Lyapunov equation, does not generally correspond to that of M. Indeed, without DVB there is typically no zeros in Θ.

We see from (16) that (M, C) satisfies DVB if and only if

$$M = -\frac{1}{2} C\Theta \tag{17}$$

where $\Theta = \Sigma^{-1}$. From this identity it is clear (since C is assumed diagonal) that under the DVB assumption, M and Θ share the same sparsity pattern. That is, we can from M directly read of the zero entries of the precision matrix for all the M-selfdecomposable distributions given by a Lévy process with $C = \mathbb{V}(Z_1)$.

Starting out with a PD matrix Θ instead, Lemma 6.3 in [Det+23] gives that $\Sigma = \Theta^{-1}$ solves the Lyapunov equation $M\Sigma + \Sigma M^T + C = 0$ if and only if

$$M = \frac{1}{2} (K - C)\, \Theta, \tag{18}$$

where K is a skew-symmetric matrix. Moreover, any M given by (18) is stable. The pair (M, C) then satisfies DVB if and only if $KC + CK = 0$ if and only if $K = 0$. Thus, for each pair (Θ, C) there is precisely one stable M, given by (17), for which (M, C) satisfies DVB and gives $\Sigma = \Theta^{-1}$ as the solution of the corresponding Lyapunov equation.

4.2. **Causal interpretation.** In the following proposition we consider the same block decomposition of M and X as in Proposition 2.3.

Proposition 4.3. *Suppose that M is stable and (M, C) satisfies drift-volatility balance. Then M_{22} is stable and*

$$\begin{aligned} \mathbb{E}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) &= \xi^{(2)} - (\Theta_{22})^{-1}\Theta_{21}(x^{(1)} - \xi^{(1)}) \\ &= \xi^{(2)} + \Sigma_{21}(\Sigma_{11})^{-1}(x^{(1)} - \xi^{(1)}) \end{aligned} \tag{19}$$

$$\mathbb{V}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) = (\Theta_{22})^{-1} = \Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}, \tag{20}$$

where $\Theta = \Sigma^{-1}$ and Σ is given by (16).

Proof. By (17) we see that $2M = -C\Theta$, whence $2M_{22} = -C_{22}\Theta_{22}$. Since Θ_{22} is a principal submatrix of a positive definite matrix, it is positive definite, and C_{22} is likewise diagonal and positive definite. As argued above, see (17) and (18), M_{22} is stable.

When DVB holds for (M, C) it holds for the pair of principal submatrices (M_{22}, C_{22}). Remark 3.3 and Proposition 4.2 then imply that

$$\Sigma^{(2)} = -\frac{1}{2}(M_{22})^{-1}C_{22} = -(-C_{22}\Theta_{22})^{-1}C_{22} = (\Theta_{22})^{-1}.$$

Since $(\Theta_{22})^{-1}$ equals the Schur complement $\Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}$, (20) follows from (8).

Note that we also have $2M_{21} = -C_{22}\Theta_{21}$, so

$$(M_{22})^{-1}M_{21} = (\Theta_{22})^{-1}(C_{22})^{-1}C_{22}\Theta_{21} = (\Theta_{22})^{-1}\Theta_{21} = -\Sigma_{21}(\Sigma_{11})^{-1}.$$

By Corollary 3.2, $a^{(2)} = -M_{21}\xi^{(1)} - M_{22}\xi^{(2)}$, and we get from (7) that

$$\begin{aligned}\mathbb{E}(X^{(2)} \mid \mathrm{do}(X^{(1)} = x^{(1)})) &= -(M_{22})^{-1}(a^{(2)} + M_{21}x^{(1)}) \\ &= \xi^{(2)} - (M_{22})^{-1}M_{21}(x^{(1)} - \xi^{(1)}) \\ &= \xi^{(2)} - (\Theta_{22})^{-1}\Theta_{21})(\xi^{(1)} - x^{(1)}) \\ &= \xi^{(2)} + \Sigma_{21}(\Sigma_{11})^{-1}(\xi^{(1)} - x^{(1)}),\end{aligned}$$

which shows (19). □

Under drift-volatility balance, the formulas (19) and (20) for the interventional mean and covariance are given entirely in terms of the observational mean vector and covariance matrix, and the formulas are the same as the observational conditional mean and covariance for the Gaussian distribution *even for non-Gaussian OSDs*. They are also the observational conditional mean and covariance more generally in the class of *elliptical distributions* [CHS81, Corollary 5].

4.3. **Rank constraints implied by drift-volatility balance.** Recall that drift-volatility balance implies

$$M = -\frac{1}{2}C\Theta. \tag{21}$$

We can therefore write the Lyapunov equation for the third order cumulant tensor under drift-volatility balance, using the Einstein summation convention, as

$$c_i(\Theta_{il}K^{ljk}) + c_j(\Theta_{jl}K^{ilk}) + c_k(\Theta_{kl}K^{ijl}) = 2d_3^v\delta_{ijkv}. \tag{22}$$

Let e_i denote the standard basis vector in $\mathbb{R}^p$ for $i = 1, \dots, p$, and define for $i, j, k \in \{1, \dots, p\}$ the p-dimensional vectors

$$v^1_{ijk} = (\Theta_{il}K^{ljk})e_i, \qquad v^2_{ijk} = (\Theta_{jl}K^{ilk})e_j, \qquad v^3_{ijk} = (\Theta_{kl}K^{ijl})e_k,$$

and

$$v_{ijk} = v^1_{ijk} + v^2_{ijk} + v^3_{ijk}. \tag{23}$$

When $i < j < k$, these vectors look as

$$v^T_{ijk} = (0, \dots, 0, \Theta_{il}K^{ljk}, 0, \dots, 0, \Theta_{jl}K^{ilk}, 0 \dots, 0, \Theta_{kl}K^{ijl}, 0, \dots, 0)$$

where $\Theta_{il}K^{ljk}$ is in the i-th entry etc. Unless $i = j = k$, (22) implies that with $c = (c_1, \ldots, c_p)^T$,

$$v_{ijk}^T c = 0.$$

If $\mathbf{V}$ denotes the matrix with p columns constructed by stacking the v_{ijk}^T-rows for $i \leq j \leq k$ not all equal, then under drift-volatility balance, $\mathbf{V}$ is rank deficient because c is non-zero. That is, $\text{rank}(\mathbf{V}) < p$.

Obviously, for the Gaussian case where $K^{ijk} = 0$, $\mathbf{V}$ is the zero matrix. We have shown (this is not included in the current draft of the manuscript) that if $d_3^v = \text{cum}_3(Z_1^v) > 0$ for $v = 1, \ldots, p$ and the directed graph induced by M is connected, then generically $\text{rank}(\mathbf{V}) = p - 1$, which under DVB allows us to identify c up to a scaling factor, and then M from Θ and c via (17). If the graph is not connected, we should consider the problem for each connectivity component, which in the extreme case of a diagonal M makes c completely unidentifiable.

We conjecture that if drift-volatility balance does not hold, then generically $\text{rank}(\mathbf{V}) = p$ whenever the graph is connected and the distribution is non-Gaussian OSD with $d_3^v = \text{cum}_3(Z_1^v) > 0$.

We propose to turn the rank deficiency constraint into a test of drift-volatility balance for non-Gaussian OSDs by computing an estimate of $\mathbf{V}$ from estimates of Θ and K, and then test if its rank is strictly less than p, e.g., via the smallest singular value.

Example 4.4. We investigate in this example with $p = 4$ the theoretical and empirical effect of drift-volatility balance on the singular values, $\lambda_1 \geq \ldots \geq \lambda_4 \geq 0$, of $\mathbf{V}$. We consider the following precision matrices parameterized by θ:

$$\Theta_\theta = \begin{pmatrix} 1 & \theta & 0 & \theta \\ \theta & 1 & \theta & 0 \\ 0 & \theta & 1 & \theta \\ \theta & 0 & \theta & 1 \end{pmatrix}. \tag{24}$$

These are positive definite (and hence valid precision matrices) for $\theta \in (-0.5, 0.5)$, and they form a submodel of the PD_G model for G the (undirected) four-cycle, cf. Figure 1. If $C = \mathbb{V}(Z_1)$ is diagonal, the corresponding M-matrices fulfilling DVB (given by (17)) have the same sparsity pattern and correspond to the left-most directed four-cycle in Figure 3.

In this example, we make $C = \mathbb{V}(Z_1) = I$ by letting the Lévy process Z have independent coordinates each being a compound Poisson process with intensity $1/4$ and jump size 2. The M-matrices that satisfy DVB are then $M_\theta = -\frac{1}{2}\Theta_\theta$.

To study the effects of drift-volatility violations, consider the skew-symmetric matrices given by

$$K_\rho = \rho \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix},$$

parametrized by $\rho \in \mathbb{R}$. Define the parametrized family of M-matrices, in accordance with (18), as

$$M_{\theta,\rho} = \frac{1}{2}(K_\rho - I)\Theta_\theta.$$

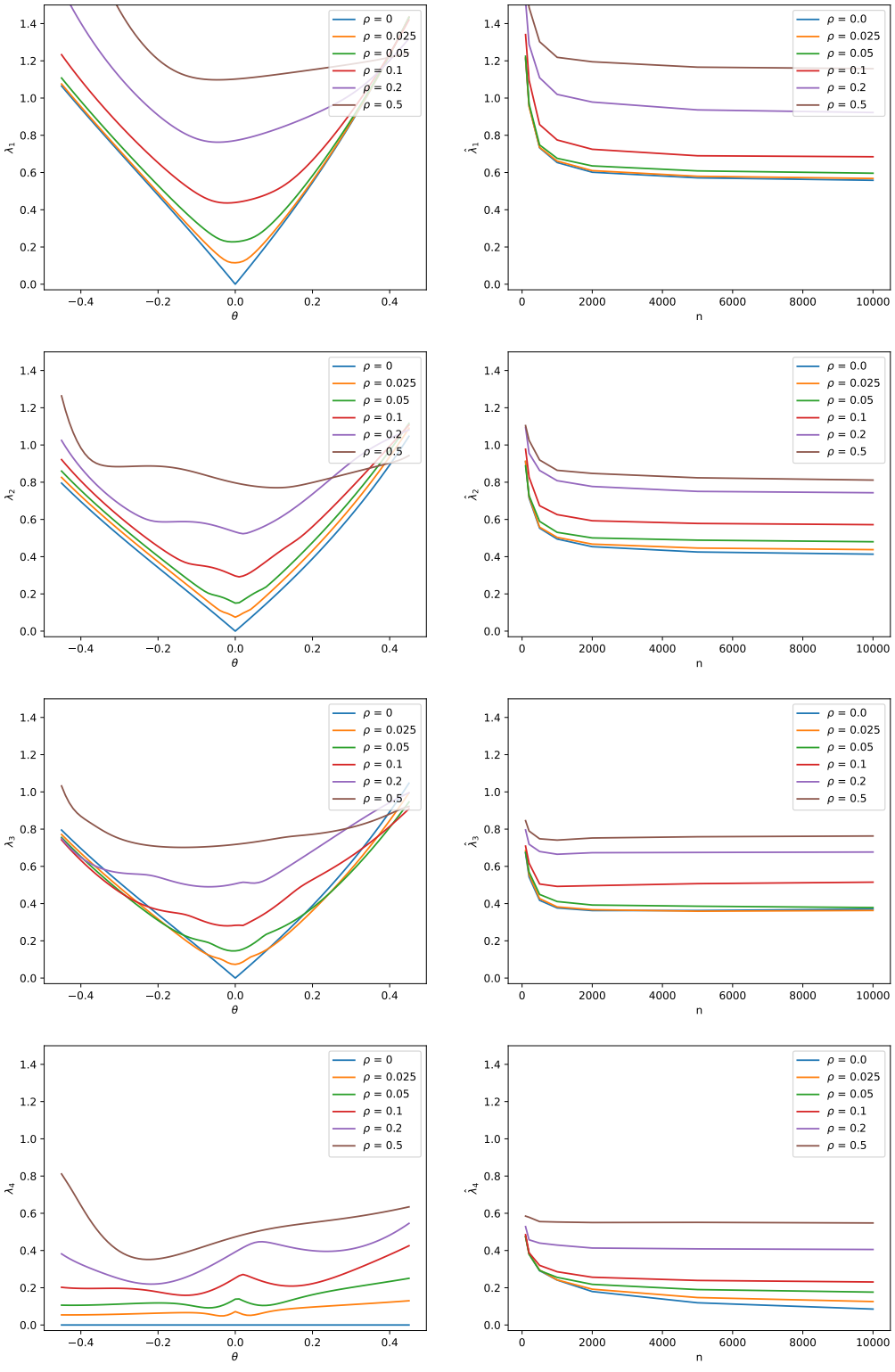


FIGURE 4. Left: Theoretical singular values λ_i of $\mathbf{V}_{\theta,\rho}$ as a function of θ (controlling the precision matrix) and for a range of skew-symmetric magnitudes ρ. Right: Corresponding trajectory of estimated singular values as a function of sample size in the case where $\theta = 0.2$.

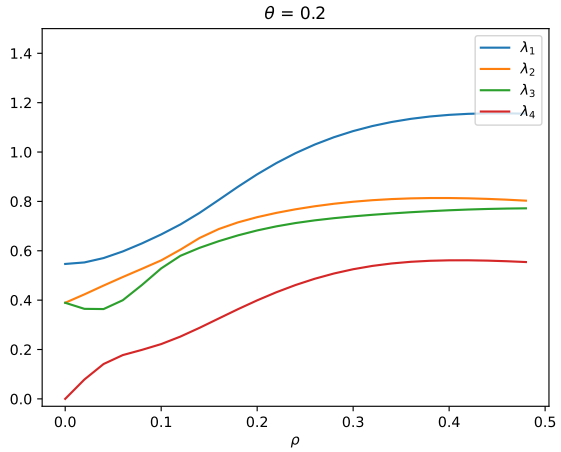


FIGURE 5. Singular values of $\mathbf{V}_{0.2,\rho}$ as a function of ρ.

The Lyapunov equation defined by the pair $(M_{\theta,\rho}, I)$ is solved by Θ_θ^{-1}, and only in the case $\rho = 0$ does the pair satisfy DVB. We can calculate the theoretical $\mathbf{V}_{\theta,\rho}$ according to (23), as well as its theoretical singular values. These are shown in the left panel of Figure 4. See also Figure 5, which shows the singular values as a function of ρ for $\theta = 0.2$. Most importantly, we see that the smallest singular value, λ_4, is equal to 0 for $\rho = 0$ when DVB is satisfied, and that λ_4 grows with ρ for all values of θ. The other singular values are strictly positive except for $\theta = \rho = 0$, in which case they are all 0 (because $\mathbf{V}_{0,0}$ is the zero-matrix). These empirical results are in accordance with our theoretical results and conjectures.

We also investigate the empirical singular values of the estimate $\hat{\mathbf{V}}_{\theta,\rho}$ calculated from a finite sample as a plug-in estimate based on $\hat{\Theta}$ and $\hat{K}$ (the empirical third order cumulant). Because the coordinates of Z are compound Poisson processes, X can be efficiently generated using the representation in (3) up to expected numerical precision of 10^{-8}.

For each θ, ρ, and n, we calculate the empirical singular values of $\hat{\mathbf{V}}_{\theta,\rho}$ and average the results over 100 trials. The right panel in Figure 4 shows the results for $\theta = 0.2$ and a range of values of ρ. We see that empirical estimates $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ are quite accurate at $n = 2000$. On the other hand, when $\rho < 0.1$, the estimate of the smallest singular value, $\hat{\lambda}_4$, is generally less accurate even for large sample sizes. Nevertheless, as can be seen from Figure 6, $\hat{\lambda}_4$ does continue to decrease with n under DVB (when $\theta = 0.2$); this does not occur when DVB is violated.

Code is available at `https://github.com/jgadams7/LinearDetailedBalance`.

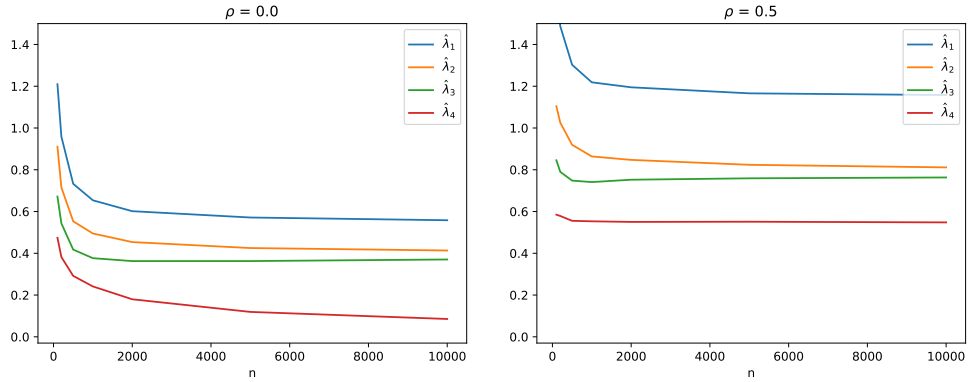


FIGURE 6. Empirical singular values of $\hat{\mathbf{V}}_{0.2,\rho}$ at two different values of ρ. Note that $\rho = 0$ corresponds to the case where drift-volatility balance is satisfied.

REFERENCES

[BGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data". In: *Journal of Machine Learning Research* 9 (2008), pp. 485–516.

[CHS81] Stamatis Cambanis, Steel Huang, and Gordon Simons. "On the theory of elliptically contoured distributions". In: *Journal of Multivariate Analysis* 11.3 (1981), pp. 368–385.

[CW00] David Roxbee Cox and Nanny Wermuth. "On the Generation of the Chordless Four-Cycle". In: *Biometrika* 87.1 (2000), pp. 206–212.

[Det+23] Philipp Dettling et al. "Identifiability in Continuous Lyapunov Models". In: *SIAM Journal on Matrix Analysis and Applications* 44.4 (2023), pp. 1799–1821.

[DM17] Mathias Drton and Marloes H. Maathuis. "Structure Learning in Graphical Modeling". In: *Annual Review of Statistics and Its Application* 4.1 (2017), pp. 365–393.

[FD11] Michael Finegold and Mathias Drton. "Robust graphical modeling of gene networks using classical and alternative t-distributions". In: *The Annals of Applied Statistics* 5.2A (2011), pp. 1057–1080.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[Fry90] Morten Frydenberg. "The Chain Graph Markov Property". In: *Scandinavian Journal of Statistics* 17.4 (1990), pp. 333–353.

[Jur82] Zbigniew J. Jurek. "An integral representation of operator-selfdecomposable random variables". In: *Bulletin de l'Académie Polonaise des Sciences. Série des Sciences Mathématiques* 30 (Jan. 1982).

[Lau96] Steffen L. Lauritzen. *Graphical models*. Vol. 17. Oxford Statistical Science Series. New York: The Clarendon Press Oxford University Press, 1996.

[LR02] Steffen L. Lauritzen and Thomas S. Richardson. "Chain graph models and their causal interpretations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 321–348.

[Mas04] Hiroki Masuda. "On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process". In: *Bernoulli* 10.1 (2004), pp. 97–120.

[Rav+11] Pradeep Ravikumar et al. "High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence". In: *Electronic Journal of Statistics* 5.none (2011), pp. 935–980.

[SH14] Alexander Sokol and Niels Richard Hansen. "Causal interpretation of stochastic differential equations". In: *Electron. J. Probab.* 19.100 (2014), pp. 1–24.

[SY84] Ken-iti Sato and Makoto Yamazato. "Operator-selfdecomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type". In: *Stochastic Processes and their Applications* 17.1 (1984), pp. 73–100.

[Urb72] Kazimierz Urbanik. "Lévy's probability measures on Euclidean spaces". eng. In: *Studia Mathematica* 44.2 (1972), pp. 119–148.

[VF11] Daniel Vogel and Roland Fried. "Elliptical graphical modelling". In: *Biometrika* 98.4 (2011), pp. 935–951.

[Wol82] Stephen J. Wolfe. "A Characterization of Certain Stochastic Integrals". In: *Stochastic Processes and their Applications (Tenth conference on stochastic processes and their applications: Montréal, Canada, 23–28 August 1981)* 12.2 (1982), pp. 117–170.

[XW21] Xiangjian Xu and Qing-Wen Wang. "On the solutions of a class of tensor equations". In: *Linear and Multilinear Algebra* (2021), pp. 1–14.

[YL07] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (Mar. 2007), pp. 19–35.

Email address: cor@math.ku.dk

Email address: ja@math.ku.dk

Email address: Niels.R.Hansen@math.ku.dk

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

4.1 Alternative Proof of the Cumulant Equations

The proof of Proposition 3.1 of [Precision] makes no assumption about the dependency structure of the p-dimensional Lévy process Z. In this section we provide an alternative proof of Proposition 3.1 in the special case where the p-dimensional Lévy process is an affine mixture of q pairwise independent Lévy processes. Specifically, we consider a p dimensional stochastic process $X = (X_1, ..., X_p)^T$ driven by a q-dimensional Lévy process (or for some of our results, merely a semimartingale) $Z = (Z_1, ..., Z_q)^T$:

$$dX_i(t) = \sum_{k=1}^{p} M_i^k X_k(t)dt + \sum_{l=1}^{q} D_i^l dZ_l(t). \tag{1}$$

Here each M_i^k (drift parameters) and D_i^l (volatility parameters) are real constants. We follow the convention of Chapter 2; superscripts index columns, and subscripts index rows.

In particular, we show that under (1), $\mathcal{C}_n$ from Proposition 3.1 has the following form:

$$\mathcal{C}_n = \sum_{l} v_l(n) \bigotimes^{k} D^l \tag{2}$$

where $\bigotimes^n$ denotes the n-way outer product and where $v_l(n)$ is a scalar that depends only on the Lévy measure of Z_l.

This representation has two advantages. First, we might be interested in the identifiability of M and $\mathcal{C}_n$ from the cumulants of X. If there is no known structure on $\mathcal{C}_n$, then the number of free parameters in $\mathcal{C}_n$ is equal to the number of cumulant equations due to Proposition 3.1. (both on the order of p^n); hence identification is impossible. However, in (2), $\mathcal{C}_n$ has only $p^2 + p$ free parameters; this fact makes identifiability of M and $\mathcal{C}_n$ much more plausible. The second advantage is that if $\mathcal{C}_n$ is identifiable (as, for example, when M is known a priori) and if no Z_l is a Brownian motion, then Kruskall's theorem entails that D is identifiable from $\mathcal{C}_k$ up to permutation and scaling of columns.

In addition to the special structure of $\mathcal{C}_n$, the calculations in this section show explicitly how the cumulant equations detailed there arise from the underlying mechanics of the Lévy-driven Ornstein-Uhlenbeck process. Many of the intermediate results require weaker assumptions than Proposition 3.1, and may be of independent interest when studying the time dynamics of Ornstein-Uhlenbeck processes.

Notation. X and Z are always understood as time-dependent processes so that their time-indices may occasionally be suppressed. A process's left limit is $X^-(t) := \lim_{s \to t, s<t} X(s)$, and its jumps are $\Delta X := X - X^-$. We follow Einstein's summation convention—indices that appear twice in a product are summed over by default—reserving the index k to index X and l to index Z. The relevant sums should be understood as preceding the entire formula. Hence (1) becomes $dX_i(t) = M_i^k X_k(t)dt + D_i^l dZ_l(t)$. We reiterate that in contrast to the usual Einstein convention, subscripts index rows and

superscripts index columns throught this thesis for consistency with Chapter 2. We write $\prod_{m\neq j}^{n} X_m := X_1...X_{j-1}X_{j+1}...X_n$ to denote the product of X_1 through X_n except X_j; similarly, when J is a set, $\prod_{m\notin J}^{n} X_m := \prod_{m\in\{1,...,n\}-J} X_m$. The covariation of two semimartingales M and N is written $[M,N]$, and the quadratic variation of M is $[M,M] =: [M]_2$. We recursively define the n-th variation of M as

$$[M]_1 := M$$
$$[M]_{n+1} := [M,[M]_n]$$

Sticklers for index counting may view expressions like $[Z_l]_n \prod_{j=1}^{n} D_j^l$ as recursively defined abbreviations for

$$[Z_l]_n \prod_{j=1}^{n} D_j^l := \left[Z_u D_n^v,\ [Z_{u'}]_{n-1} \prod_{j=1}^{n-1} D_j^{v'} \right] \delta_l^{u,u'} \delta_{v,v'}^l$$

where δ is the Kronecker delta.

To begin, we prove an interesting relation between the n-th variation of a semimartingale and its jumps. This allows us to conclude that the n-th variation has finite variation, so that we may integrate with respect to it.

Proposition 1. *If Z is a semimartingale, then $[Z]_n(t) = \sum_{s\leq t} \Delta Z(s)^n$ for $n \geq 3$. Moreover, $[Z]_n$ has finite variation for all $n \geq 2$.*

Proof. All proofs are in Section 4.1.1 □

Using Ito's formula for discontinuous processes and mathematical induction, we can show that:

Proposition 2. *If X is a semimartingale, then for indices $i_1, ..., i_n$:*

$$d\prod_{j=1}^{n} X_{i_j} = \sum_{j=1}^{n} \left(\prod_{m\neq j}^{n} X_{i_m}^- \right) dX_{i_j} + \left(\prod_{m=j+1}^{n} X_{i_m}^- \right) d\left[X_{i_j}, \prod_{m=1}^{j-1} X_{i_m} \right].$$

It is desirable to rewrite the stochastic integrals with respect to covariations of X as stochastic integrals with respect to something more basic. To this end, we leverage (1), as well as the facts that time integrals of real-valued functions are absolutely continuous, that Z is a semimartingale, and that the covariation is bilinear.

Proposition 3. *If X is given as in (1) and $Z_1, ..., Z_q$ are pairwise independent semi-*

martingales, then for indices $i_1, ..., i_n$*:*

$$d\prod_{j=1}^{n} X_{i_j} = \sum_{j=1}^{n} \left(\prod_{m \neq j}^{n} X_{i_m}^{-} \right) M_{i_j}^{k} X_k dt + \sum_{J \subset \{1,...,n\}} d[Z_l]_{n-|J|} \prod_{m \notin J}^{n} D_{i_m}^{l} \prod_{m \in J} X_{i_m}^{-}$$

To get moment equations out of Proposition 3, we need to integrate and then take expectations. It can be seen that all resulting integrals are of one of two forms:

1. $\mathbb{E} \int \text{poly}(X^-) X_k dt$
2. $\mathbb{E} \int \text{poly}(X^-) d[Z_l]_m$ for $m \geq 2$

While the first is potentially estimable under mild assumptions, the second is harder to estimate without strong assumptions due to its dependence on the latent processes via $[Z_l]_m$. One sufficiently strong assumption is that Z is a Lévy process.

Proposition 4. *If* X *is given as in* (1) *and* $Z_1, ..., Z_q$ *are pairwise independent Lévy processes, then for indices* $i_1, ..., i_n$*:*

$$\mathbb{E}\prod_{j=1}^{n} X_{i_j} = \sum_{j=1}^{n} M_{i_j}^{k} \mathbb{E} \int \left(\prod_{m \neq j}^{n} X_{i_m} \right) X_k dt + \sum_{J \subset \{1,...,n\}} \left(\prod_{m \notin J}^{n} D_{i_m}^{l} \right) \mathbb{E}[Z_l]_{n-|J|}(1) \mathbb{E} \int \prod_{m \in J} X_{i_m} dt + \mathbb{E} \prod_{j=1}^{n} X_{i_j}(0)$$

Notice that the quantities $\mathbb{E} \int \left(\prod_{m \in J} X_{i_m} \right) dt$ are estimable from data, and that the quantities $\mathbb{E}([Z_l]_2(1))$ may be set to 1 without loss of generality by rescaling D and Z accordingly.

Proposition 4 is a nearly immediate consequence of the following fact:

Proposition 5. $[Z]_n$ *is a Lévy process whenever* Z *is a Lévy process.*

Proposition 4 describes the moments of $X(T)$ in terms of the time dynamics up to time T. In the case where T is large and X has reached a steady-state distribution, the individual time dynamics become negligible and we arrive at a much simpler moment equation.

Proposition 6. *In addition to the assumptions of Proposition 4, suppose that X is stationary. Then for any time $t \geq 0$,*

$$0 = \sum_{j=1}^{n} M_{i_j}^k \mathbb{E}\left(X_k \prod_{m \neq j}^{n} X_{i_m} \right) + \sum_{J \subset \{1,\dots,n\}} \left(\prod_{m \notin J}^{n} D_{i_m}^l \right) v_l(n - |J|) \mathbb{E} \prod_{m \in J} X_{i_m}$$

where $v_l(i) := \mathbb{E}[Z_l]_i(t = 1)$

By rewriting Proposition 6 in terms of cumulants, it is possible to absorb much of the second term (which is polynomial in D, lower moments of X, and variations of Z) into the first term (which is linear in M). An induction argument shows that:

Proposition 7. *Under the conditions of Proposition 6, it is the case that*

$$0 = \sum_{j=1}^{n} M_{i_j}^k \mathbb{K}\left(X_k \prod_{m \neq j}^{n} X_{i_m} \right) + v_l(n) \prod_{m=1}^{n} D_{i_m}^l$$

where $v_l(i) := \mathbb{E}[Z_l]_i(t = 1)$

where $\mathcal{K}(n)_{i_1,\dots,i_n} := \mathbb{K}(\prod_{j=1}^{n} X_{i_j})$ is the n-th cumulant of $\{X_{i_j}\}_{j=1}^{n}$.

The clear advantage of the cumulant equations in Proposition 7 over the moment equations in Proposition 6 is that the former depend only on a single order cumulant of X. By contrast, the moment equation involves a complicated interplay between D and all lower-order moments of X.

In tensor form, the first term of Proposition 7 can be written

$$\left(\sum_{j=1}^{n} \mathcal{K}(n) \times_j M^T \right)_{i_1,\dots,i_n}.$$

This means that Proposition 7 has the form of a continuous Lyapunov equation,

$$0 = \sum_{j=1}^{n} \mathcal{K}(n) \times_j M^T + \mathcal{C}(n).$$

Here,

$$\mathcal{C}(n) = \sum_{l} v_l(n) \bigotimes^{n} D^l = \mathcal{V}(n) \times_1 D^T \dots \times_n D^T$$

where $\mathcal{V}(n)_{i_1,\dots,i_n} := v_l(n) \delta^l_{i_1,\dots,i_j}$ is an n-order p^n-dimensional tensor with $\mathbb{E}[Z]_n(1)$ along the main diagonal and zero elsewhere.

4.1.1 Proofs

4.1.1.1 Proof of Proposition 1

Proof. For $n = 3$,

$$[Z,[Z,Z]] = \int \Delta Z d[Z,Z] = \sum_{s\leq t} \Delta Z \Delta[Z,Z] = \sum_{s\leq t} \Delta Z (\Delta Z)^2$$

where the first and second equalities follow because $[Z,Z]$ has finite variation. To see that $[Z]_3$ has finite variation, notice

$$\begin{aligned}
\sum_{s\leq t} |\Delta Z|^3 &= \sum_{s\leq t:|\Delta Z|<1} |\Delta Z|^3 + \sum_{s\leq t:|\Delta Z|\geq 1} |\Delta Z|^3 \\
&\leq \sum_{s\leq t:|\Delta Z|<1} |\Delta Z|^2 + \sum_{s\leq t:|\Delta Z|\geq 1} |\Delta Z|^4 \\
&\leq \sum_{s\leq t} |\Delta Z|^2 + \sum_{s\leq t} |\Delta Z|^4 \\
&= [Z,Z] + \sum_{s\leq t} (\Delta[Z,Z])^2 \\
&= [Z,Z] + [[Z,Z],[Z,Z]] < \infty
\end{aligned}$$

For $n > 3$,

$$\begin{aligned}
[Z,[Z]_{n-1}] =& [Z, \int (\Delta Z)^{n-3} d[Z,Z]] \\
=& \int (\Delta Z)^{n-3} d[Z,[Z,Z]] \\
=& \int (\Delta Z)^{n-2} d[Z,Z] \\
=& \sum_{s\leq t} (\Delta Z)^n
\end{aligned}$$

where the first and third equalities follow by induction. The argument that $[Z]_n$ has finite variation is similar to the case when $n = 3$. When n is even,

$$\begin{aligned}
\sum_{s\leq t} (\Delta Z)^n &= \sum_{s\leq t} (\Delta [Z]_{n/2})^2 \\
&= \left[[Z]_{n/2},[Z]_{n/2}\right]
\end{aligned}$$

and when n is odd,

$$\begin{aligned}\sum_{s\leq t}(\Delta Z)^n &\leq \sum_{s\leq t}(\Delta Z)^{n-1} + \sum_{s\leq t}(\Delta Z)^{n+1} \\ &= [\Delta Z]_{n-1} + \left[[Z]_{(n+1)/2}, [Z]_{(n+1)/2}\right]\end{aligned}$$

In either case, the result is finite by induction. □

4.1.1.2 Proof of Proposition 2

Proof. Proceed by induction. For $n = 1$, the result holds because $\prod_{j\in\emptyset} X_j = 1$ and $[X_j, 1] = 0$.

For the inductive step, use Ito's formula for discontinuous processes:

$$\begin{aligned}d\prod_{j=1}^{n} X_{i_j} =& d\left(X_{i_n}\prod_{j=1}^{n-1} X_{i_j}\right) \\ =& X^-_{i_n} d\prod_{j=1}^{n-1} X_{i_j} + \left(\prod_{j=1}^{n-1} X^-_{i_j}\right) dX_{i_n} + d\left[X_{i_n}, \prod_{j=1}^{n-1} X_{i_j}\right] \\ &+ \Delta\prod_{j=1}^{n} X_{i_j} - X^-_{i_n}\Delta\prod_{j=1}^{n-1} X_{i_j} - \left(\prod_{j=1}^{n-1} X^-_{i_j}\right)\Delta X_{i_n} - \Delta X_{i_n}\Delta\prod_{j=1}^{n-1} X_{i_j} \\ =& X^-_{i_n} d\prod_{j=1}^{n-1} X_{i_j} + \left(\prod_{j=1}^{n-1} X^-_{i_j}\right) dX_{i_n} + d\left[X_{i_n}, \prod_{j=1}^{n-1} X_{i_j}\right] + 0 \\ =& X^-_{i_n}\sum_{j=1}^{n-1}\left(\prod_{m\neq j}^{n-1} X^-_{i_m}\right) dX_{i_j} + \left(\prod_{m=j+1}^{n-1} X^-_{i_m}\right) d\left[X_{i_j}, \prod_{m=1}^{j-1} X_{i_m}\right] \\ &+ \left(\prod_{j=1}^{n-1} X^-_{i_j}\right) dX_{i_n} + d\left[X_{i_n}, \prod_{j=1}^{n-1} X_{i_j}\right] \\ =& \sum_{j=1}^{n}\left(\prod_{m\neq j}^{n} X^-_{i_j}\right) dX_{i_j} + \left(\prod_{m=j+1}^{n} X^-_{i_j}\right) d\left[X_{i_j}, \prod_{m=1}^{j-1} X_{i_m}\right]\end{aligned}$$

where the fourth equality invokes the inductive hypothesis. □

4.1.1.3 Proof of Proposition 3

Our proof of Proposition 3 utilizes the following technical lemma.

Lemma 1. *Let*

$$\psi_n := \sum_{j=1}^{n} \left(\prod_{m=j+1}^{n} X_{i_m}^-(t) \right) \left(D_{i_j}^l \sum_{J \subseteq \{1,\ldots,j-1\}} d[Z_l]_{j-|J|+c} \prod_{m \notin J}^{j-1} D_{i_m}^l \prod_{m \in J} X_{i_m}^- \right),$$

where c is any whole number and $Z_1, \ldots, Z_q$ are semimartingales. Then

$$\psi_n = \sum_{J \subset \{1,\ldots,n\}} d[Z_l]_{n-|J|+c} \prod_{m \notin J}^{n} D_{i_m}^l \prod_{m \in J} X_{i_m}^-$$

for all $n \geq 1$

Proof. We proceed by induction. For $n = 1$, we recall that $\prod_{m \in \emptyset} X_m = 1$ so that both expressions equal $d[Z_l]_{1+c} D_{i_1}^l$ as desired. For $n > 1$, separate the sum into $j = n$ and $1 \leq j < n$ to obtain:

$$\begin{aligned}
\psi_n =& X_{i_n}^- \sum_{j=1}^{n-1} \left(\prod_{m=j+1}^{n-1} X_{i_m}^-(t) \right) \left(D_{i_j}^l \sum_{J \subseteq \{1,\ldots,j-1\}} d[Z_l]_{j-|J|+c} \prod_{m \notin J}^{j-1} D_{i_m}^l \prod_{m \in J} X_{i_m}^- \right) \\
&+ D_{i_n}^l \sum_{J \subseteq \{1,\ldots,n-1\}} d[Z_l]_{n-|J|+c} \prod_{m \notin J}^{n-1} D_{i_m}^l \prod_{m \in J} X_{i_m}^- \\
=& (X_{i_n}^- + D_{i_n}^l) \sum_{J \subset \{1,\ldots,n-1\}} d[Z_l]_{n-|J|+c} \prod_{m \notin J}^{n-1} D_{i_m}^l \prod_{m \in J} X_{i_m}^- \\
&+ D_{i_n}^l \sum_{J = \{1,\ldots,n-1\}} d[Z_l]_{n-|J|+c} \prod_{m \notin J}^{n-1} D_{i_m}^l \prod_{m \in J} X_{i_m}^-
\end{aligned}$$

where the second equality is obtained by applying the inductive hypothesis to the first term and separating the second term into $J \subset \{1, \ldots, n-1\}$ and $J = \{1, \ldots, n-1\}$. The result follows by comparing terms. □

We now prove Proposition 3.

Proof. First we observe that if Z is a semimartingale, then X is also a semimartingale if it evolves according to (1), so that the conditions of Proposition 2 are satisfied.

Next we argue that

$$d\left[X_{i_{n+1}}, \prod_{m=1}^{n} X_{i_m} \right] = D_{i_{n+1}}^l \left(\sum_{J \subset \{1,\ldots,n\}} d[Z_l]_{n-|J|+1} \prod_{m \notin J} D_{i_m}^l \prod_{m \in J} X_{i_m}^- \right), \quad n > 0$$

by induction on n. For $n = 1$,

$$\begin{aligned} d[X_{i_2}, X_{i_1}] =& d[D^l_{i_2} Z_l, Z_{l'} D^{l'}_{i_1}] \\ =& D^l_{i_2} d[Z_l]_2 D^l_{i_1} \\ =& D^l_{i_2} \left(\sum_{J \subset \{1\}} d[Z_l]_{2-|J|} \prod_{m \notin J}^{1} D^l_{i_m} \prod_{m \in J} X^-_{i_m} \right) \end{aligned}$$

where the first equality follows from (1), the bilinearity of the covariation, and the fact that $\int M^k_i X_k dt$ is absolutely continuous and that Z_l is a martingale; and the second follows from the fact that Z_l are pairwise independent.

For $n > 1$,

$$\begin{aligned} \left[X_{i_{n+1}}, \prod_{m=1}^{n} X_{i_m} \right] =& \left[D^l_{i_{n+1}} Z_l, \int \left(\prod_{m \neq j}^{n} X^-_{i_m}(t) \right) D^l_{i_j} dZ_l \right] \\ &+ \left[D^l_{i_{n+1}} Z_l, \int \left(\prod_{m=j+1}^{n} X^-_{i_m}(t) \right) d \left[X_{i_j}, \prod_{m=1}^{j-1} X_{i_m} \right] \right] \\ =& \left[D^l_{i_{n+1}} Z_l, \int \left(\prod_{m=j+1}^{n} X^-_{i_m}(t) \right) \left(D^l_{i_j} d[Z]_l \prod_{m=1}^{j-1} X^-_{i_m} \right) \right] \\ &+ \left[D^l_{i_{n+1}} Z_l, \int \left(\prod_{m=j+1}^{n} X^-_{i_m}(t) \right) \left(D^l_{i_j} \sum_{J \subset \{1,\dots,j-1\}} d[Z_l]_{j-|J|} \prod_{m \notin J}^{j-1} D^l_{i_m} \prod_{m \in J} X^-_{i_m} \right) \right] \\ =& \left[D^l_{i_{n+1}} Z_l, \int \left(\prod_{m=j+1}^{n} X^-_{i_m}(t) \right) \left(D^l_{i_j} \sum_{J \subseteq \{1,\dots,j-1\}} d[Z_l]_{j-|J|} \prod_{m \notin J}^{j-1} D^l_{i_m} \prod_{m \in J} X^-_{i_m} \right) \right] \\ =& \int D^l_{i_{n+1}} \left(\prod_{m=j+1}^{n} X^-_{i_m}(t) \right) \left(D^l_{i_j} \sum_{J \subseteq \{1,\dots,j-1\}} d[Z_l]_{j-|J|+1} \prod_{m \notin J}^{j-1} D^l_{i_m} \prod_{m \in J} X^-_{i_m} \right) \\ =& \int D^l_{i_{n+1}} \sum_{J \subset \{1,\dots,n\}} d[Z_l]_{n-|J|+1} \prod_{m \notin J}^{n} D^l_{i_m} \prod_{m \in J} X^-_{i_m} \end{aligned}$$

where the first equality follows from (1) and Proposition 2, the bilinearity of covariation, and the fact that terms like $\int \text{poly}(X^-) X_k ds$ are always absolutely continuous while terms like $\int \text{poly}(X^-) dZ$ are semimartingales when Z is a semimartingale; the second equality applies the inductive hypothesis to the second term; the fourth equality uses the fact that $[Z_l, \int \text{poly}(X^-) d[Z_{l'}]_k] = \int \text{poly}(X^-) d[Z_l, [Z_{l'}]_k]$ and that independent semimartingales have covariation zero; and the fifth equality uses Lemma 1 with $c = 1$.

Substituting into Proposition 2 gives

$$
\begin{aligned}
d\prod_{j=1}^{n} X_{i_j}(t) =& \sum_{j=1}^{n}\left(\prod_{m\neq j}^{n} X_{i_m}^{-}(t)\right) dX_{i_j} \\
&+ \left(\prod_{m=j+1}^{n} X_{i_m}^{-}(t)\right)\left(D_{i_j}^{l}\sum_{J\subset\{1,\ldots,j-1\}} d[Z_l]_{j-|J|}\prod_{m\notin J}^{j-1} D_{i_m}^{l}\prod_{m\in J} X_{i_m}^{-}\right) \\
=& \sum_{j=1}^{n}\left(\prod_{m\neq j}^{n} X_{i_m}^{-}(t)\right)(M_{i_j}^{k}X_k dt + D_{i_j}^{l} dZ_l) \\
&+ \left(\prod_{m=j+1}^{n} X_{i_m}^{-}(t)\right)\left(D_{i_j}^{l}\sum_{J\subseteq\{1,\ldots,j-1\}} d[Z_l]_{j-|J|}\prod_{m\notin J}^{j-1} D_{i_m}^{l}\prod_{m\in J} X_{i_m}^{-}\right) \\
&- \left(\prod_{m=j+1}^{n} X_{i_m}^{-}(t)\right)\left(D_{i_j}^{l} d[Z_l]_1\prod_{m=1}^{j-1} X_{i_m}^{-}\right)
\end{aligned}
$$

where the second equality substitutes (1) into the first term and adds the third term to the second term. Recalling that $[Z_l]_1 := Z_l$ and applying Lemma 1 to the second term with $c = 0$ completes the proof. □

4.1.1.4 Proof of Proposition 4

Proof. Since the conditions of Proposition 3 are satisfied, take the expectation after the integral to obtain

$$
\begin{aligned}
\mathbb{E}\prod_{j=1}^{n} X_{i_j} =& \sum_{j=1}^{n} M_{i_j}^{k}\mathbb{E}\int\left(\prod_{m\neq j}^{n} X_{i_m}^{-}\right) X_k dt \\
&+ \sum_{J\subset\{1,\ldots,n\}}\left(\prod_{m\notin J}^{n} D_{i_m}^{l}\right)\mathbb{E}\int\prod_{m\in J} X_{i_m}^{-} d[Z_l]_{n-|J|} \\
&+ \mathbb{E}\prod_{j=1}^{n} X_{i_j}(0)
\end{aligned}
$$

where the last term on the right is the integration constant from Ito's formula. Since $X^{-} = X$ a.e. a.s., the minus can be dropped from the first term. The second term is the expectation of the integral of a predictable process with respect to a Lévy process—see Proposition 5. As such,

$$
\mathbb{E}\int\prod_{m\in J} X_{i_m}^{-} d[Z_l]_{n-|J|} = \mathbb{E}\int\prod_{m\in J} X_{i_m}^{-}\lambda_l(n-|J|)dt
$$

, where $\lambda_l(j) := \mathbb{E}[Z_l]_j(t=1)$. □

4.1.1.5 Proof of Proposition 5

Proof. Clearly $[Z]_2(t) = bt + \sum_{s\leq t}(\Delta Z(s))^2$ is continuous in probability with independent and stationary increments whenever Z is. The same is true of $\sum(\Delta Z)^n$, which by Proposition 1 is equal to $[Z]_n$ for $n > 2$. □

4.1.1.6 Proof of Proposition 6

Proof.

$$\begin{aligned}\mathbb{E}\int_0^s \text{poly}(X(s))dt &= \int_0^s \mathbb{E}\text{poly}(X(s))dt \\ &= \int_0^s \mathbb{E}\text{poly}(X(0))dt \\ &= s\mathbb{E}\text{poly}(X(0))\end{aligned}$$

where the second equation follows from stationarity. Hence

$$\begin{aligned}\mathbb{E}\prod_{j=1}^n X_{i_j}(s) =& s\sum_{j=1}^n M_{i_j}^k \mathbb{E}\left(X_k(0)\prod_{m\neq j}^n X_{i_m}(0)\right) \\ &+ s\sum_{J\subset\{1,\dots,n\}}\left(\prod_{m\notin J}^n D_{i_m}^l\right)\mathbb{E}[Z_l]_{n-|J|}(1)\mathbb{E}\prod_{m\in J} X_{i_m}(0) \\ &+ \mathbb{E}\prod_{j=1}^n X_{i_j}(0)\end{aligned}$$

by Proposition 4 and the above calculation. The left hand side cancels with the third term on the right hand side due to stationarity. Dividing through by s gives the desired result for $t = 0$; again invoking stationarity gives the result for all $t \geq 0$. □

4.1.1.7 Proof of Proposition 7

Proof. For $n = 1$, we have

$$\begin{aligned}0 &= M_i^k \mathbb{E}X_k + D_i^l v_l(1) \\ &= M_i^k \mathbb{K}X_k + D_i^l v_l(1)\end{aligned}$$

For $n > 1$, the second sum is

$$\begin{aligned}
\psi(i_1,...,i_n) := &\sum_{J\subset\{1,...,n\}} \left(v_l(n-|J|) \prod_{m\notin J}^{n} D^l_{i_m} \right) \mathbb{E} \prod_{m\in J} X_{i_m} \\
=&v_l(n) \prod_{j=1}^{n} D^l_{i_j} + \sum_{\substack{J\subset\{1,...,n\} \\ J\neq\emptyset}} \left\{ -\sum_{j\notin J}^{n} M^k_{i_j} \mathbb{K}\left(X_k \prod_{m\notin J\cup\{j\}}^{n} X_{i_m} \right) \right\} \mathbb{E} \prod_{m\in J} X_{i_m} \\
=&v_l(n) \prod_{j=1}^{n} D^l_{i_j} - \sum_{j=1}^{n} M^k_{i_j} \sum_{\substack{J\subseteq\{1,...,n\}-\{j\} \\ J\neq\emptyset}} \mathbb{K}\left(X_k \prod_{m\notin J\cup\{j\}}^{n} X_{i_m} \right) \mathbb{E} \prod_{m\in J} X_{i_m} \\
=&v_l(n) \prod_{j=1}^{n} D^l_{i_j} - \sum_{j=1}^{n} M^k_{i_j} \sum_{\substack{J\subseteq\{1,...,n\}-\{j\} \\ J\neq\emptyset}} \mathbb{K}\left(X_k \prod_{m\notin J\cup\{j\}}^{n} X_{i_m} \right) \sum_{\sigma\in\Psi(J)} \prod_{H\in\sigma} \mathbb{K}\left(\prod_{h\in H} X_h \right) \\
=&v_l(n) \prod_{j=1}^{n} D^l_{i_j} - \sum_{j=1}^{n} M^k_{i_j} \left\{ \mathbb{E}\left(X_k \prod_{m\neq j}^{n} X_{i_m} \right) - \mathbb{K}\left(X_k \prod_{m\neq j}^{n} X_{i_m} \right) \right\}
\end{aligned}$$

where the second line splits the sum into $|J| = 0$ and $|J| > 0$ and then applies the inductive hypothesis to the latter; the third line follows by comparing terms; the fourth line uses the fact that the raw moment of X equals the sum of the products of the cumulants of the partitions of X (here $\Psi(J)$ refers to the partitions of J); and the fifth line again uses this fact. Therefore

$$\begin{aligned}
0 =&\sum_{j=1}^{n} M^k_{i_j} \mathbb{E}\left(X_k \prod_{m\neq j}^{n} X_{i_m} \right) + \psi(i_1,...,i_n) \\
=&v_l(n) \prod_{j=1}^{n} D^l_{i_j} + \sum_{j=1}^{n} M^k_{i_j} \mathbb{K}\left(X_k \prod_{m\neq j}^{n} X_{i_m} \right)
\end{aligned}$$

as desired. □

Bibliography

J. Adams and N. R. Hansen. Substitute adjustment via recovery of latent variables. *arXiv preprint arXiv:2403.00202*, 2024.

J. Adams, N. Hansen, and K. Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34:22822–22833, 2021.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020.

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36: 287–314, 1994.

A. D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486. PMLR, 2019.

F. M. Fisher. A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pages 73–92, 1970.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. Discovering causal structure: artificial intelligence, philosophy of science and statistical modeling. 1986.

J. Grimmer, D. Knox, and B. Stewart. Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *Journal of Machine Learning Research*, 24(182):1–70, 2023.

B. Guo, J. Nie, and Z. Yang. Learning diagonal Gaussian mixture models and incomplete tensor decompositions. *Vietnam Journal of Mathematics*, 50:421–446, 2022.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen. Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528):1611–1615, 2019.

E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen. Counterexamples to "the blessings of multiple causes" by wang and blei. *arXiv:2001.06555*, 2020.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 154–162, 2013.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

J. Qiao, R. Cai, K. Zhang, Z. Zhang, and Z. Hao. Causal discovery with confounding cascade nonlinear additive noise models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(6):1–28, 2021.

C. O. Recke, J. Adams, and N. R. Hansen. Non-Gaussian graphical precision models. 2024.

J. A. Rhodes. A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and its Applications*, 432(7):1818–1824, 2010.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Y. Wang and D. M. Blei. Towards clarifying the theory of the deconfounder. *arXiv:2003.04948*, 2020.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. pages 647–655. AUAI Press, 2009.