Journal of Machine Learning Research 26 (2025) 1-44 Submitted 2/24; Revised 8/25; Published 11/25

Substitute Adjustment via Recovery of Latent Variables

Jeffrey Adams JA@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
Copenhagen, 2100, Denmark

Niels Richard Hansen NIELS.R.HANSEN@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
Copenhagen, 2100, Denmark

Editor: David Sontag

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 and shows that it can be a viable method for adjusted regression when the recovery error is small. Most importantly, we present clear assumptions about the data generating distribution that allow us to control the recovery error and ultimately the estimation error that can be attributed to the use of substitutes.

Keywords: adjusted regression, causality, deconfounder, latent variables, mixture models

1. Introduction

The deconfounder was proposed by Wang and 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 on the paper by Wang and 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 and Blei (2019) are sufficient

to claim that the deconfounder estimates a causal parameter. Though an amendment by Wang and 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). Similarly, surrogate variable adjustment (Leek and 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.

We find that 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 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., 2024).

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 difference

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

is a measure of how much a SNP value of x at site i adds to the prediction of the phenotype outcome on top of the values $\mathbf{X}_{-i}$ of all other SNP sites. The regression function (1) thus quantifies the expected added predictive value of the SNP at site i. A causal interpretation might be justified if the SNP is located in a gene that is causal for the phenotypic outcome, but any added predictive value of a single SNP is of interest when constructing polygenic risk scores—irrespective of causal interpretations. 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 $\mathbb{E}\left[Y \mid X_i = x; \mathbf{X}_{-i} = \mathbf{x}\right]$ 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 observed 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:

1. *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.

2. *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 and 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 Kallus

et al. (2018), who give theoretical results on the estimation error due to recovery of latent confounders from noisy proxy observations. Their control of the estimation error in terms of the recovery error of the column space of the confounders resembles our Theorem 7 and its proof, but it is based on a low rank matrix factorization framework rather than the finite mixture models we consider.

Related approaches by Ćevid et al. (2020) and Guo et al. (2022b) are based on estimators within a linear modeling 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. Linearity and sparsity assumptions play an important role for their methods and analysis but not for our results. Additional approaches, that do not attempt explicit recovery of latent variables, include proxy and auxiliary variable methods as considered by Louizos et al. (2017); Miao et al. (2018, 2023); Tchetgen et al. (2024). Compared with our framework, these methods deal with unobserved confounding without asymptotic exact recovery but under particular distributional assumptions.

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. We analyze instead the asymptotic behavior 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

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 $\mathbf{X}_{-i}$. In this section we present a rigorous model specification together with the general model assumptions and substitute adjustment algorithms.

2.1 The General Model

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*

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

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 our target χ^i_x, let alone the full Markov kernel $P^i_{x,\mathbf{x}}$, from the distribution of $(\mathbf{X}, Y)$, which we will not pursue here, but see Remark 2.

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.

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

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

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

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

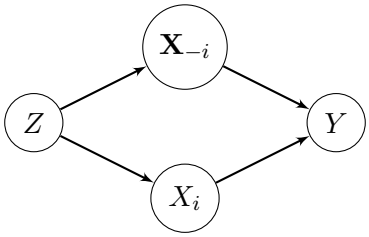

Figure 1: Directed Acyclic Graph (DAG) representing the joint distribution of $(X_i, \mathbf{X}_{-i}, Z, Y)$. The variable Z blocks the backdoor from X_i to Y.

Assumption 2(1) implies that $\mathbf{X}_{-i} \perp\!\!\!\perp X_i \mid Z$ and Assumption 2(2) implies that Z is a function of $\mathbf{X}_{-i}$. This means that Z is a *balancing score* for the "treatment" X_i and "covariates" $\mathbf{X}_{-i}$ simultaneously for all i, see Rosenbaum and Rubin (1983). Their Theorem 3 shows, within a potential outcomes framework and for binary treatments, that strong ignorability given $\mathbf{X}_{-i}$ implies strong ignorability given the balancing score Z. Consequently, under strong ignorability given $\mathbf{X}_{-i}$ the average treatment effect can be obtained by adjusting only for the balancing score Z. The same reasons that make Z a balancing score make the joint distribution of $(X_i, \mathbf{X}_{-i}, Z, Y)$ Markov w.r.t. to the graph in Figure 1. Proposition 2 is thus a variant of 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).

Even though balancing scores and the backdoor criterion are well known, we include a proof of Proposition 2 in Appendix A for the following reasons. First, Proposition 2 does not involve causal assumptions about the model, and we want to clarify that the mathematical result is agnostic to such assumptions. Second, our proof is for real valued variables X_i—not only binary or discrete variables—and it does not rely on positivity assumptions. Third, we do not need other regularity assumptions either, specifically we do not require that the conditional distributions have densities w.r.t. a fixed measure, which is an unreasonable assumption for our infinite variable model. However, to be clear, the proof of Proposition 2 simply shows that we can replace the adjustment variables in an adjusted regression by a balancing score, and that Assumption 2 makes Z a balancing score for $(X_i, \mathbf{X}_{-i})$ for all i.

To illuminate Assumptions 1 and 2—as well as to illustrate the implications of Proposition 2—we give two examples below. They show that the assumptions can be fulfilled, and they give explicit examples of the Markov kernels $Q^i_{x,z}$ and the conditional expectations

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

Example 1 Suppose $\mathbb{E}[|X_i|] \leq C$ 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

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

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 2 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 3 (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)*.*

Remark 2 As noted in Remark 1, identification of χ^i_x from the observational distribution requires additional regularity assumptions. Algorithm 1 below works if we can identify (and consistently estimate) the regression function $(x, z) \mapsto b^i_x(z)$. This can be done on the support of the distribution of (X_i, Z) if the regression function is suitably regular. We avoid a precise general statement since it will depend on the specific nature of the set E. Assumptions 3 or 4 below include explicit positivity conditions that allow for identification of an assumption-lean target parameter when E is finite.

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 $\widehat{\chi}^{i,\mathrm{sub}}_x$.

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.

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

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

7 **end**

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

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 $\widehat{\chi}^{i,\text{sub}}_x$. Thus the estimate is directly estimating $\chi^{i,\text{sub}}_x = \mathbb{E}\left[\mathbb{E}\left[Y \mid X_i = x; \hat{Z}\right] \,\middle|\, \hat{f}^p\right]$, 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, $\chi^{i,\text{sub}}_x \to \chi^i_x$ 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 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 data sets 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 and 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) Suppose $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}[X_i^2] < \infty$ and $\mathbb{E}\left[\text{Var}\left[X_i \mid Z\right]\right] > 0$.

Definition 4 (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.

Proposition 5 *Under Assumptions 1, 2 and 3, and with $b^i_x(z)$ given as in Definition 3, and β_i given as in Definition 4,*

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

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

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

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 $\mu_i(z) = \mathbb{E}\left[X_i \mid Z = z\right]$ and $g(z) = \mathbb{E}\left[Y \mid Z = z\right]$;

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

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 5 in Appendix A.1 for completeness. The arguments are essentially as given by Vansteelandt and Dukes (2022).

Remark 3 If $b^i_x(z) = \beta_i'(z)x + \Gamma_{-i}(z)$ it follows from Proposition 2 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 4 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 5 and the remarks above as follows:

1. *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.

2. *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.

3. *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 $0 < \sigma_{\min}^2 \leq \sigma_{\max}^2 < \infty$ that bound the conditional variances;

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

   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 $(x_{i,1}, \hat{z}_1, y_1), \dots (x_{i,n}, \hat{z}_n, y_n)$. This may be relevant in practice, but it is also used in the proof of Theorem 7. The corresponding oracle estimator, $\hat{\beta}_i$, is similarly an OLS estimator.

Algorithm 3: Assumption Lean Substitute Adjustment w. Mixtures

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

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 $j \in \{1, \dots, p\}$ and $z \in E$.

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

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

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

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

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 6 *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 5 The proof of Proposition 6 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 6 The arguments leading to Proposition 6 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 16 and Corollary 17 in Appendix B, which gives necessary and sufficient conditions for recovery in the Gaussian mixture model.

3.2 Bounding the 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 $(x_{i,1}, \hat{z}_1, z_1, y_1), \ldots (x_{i,n}, \hat{z}_n, z_n, y_n)$, which include the otherwise unobserved z_k-s as well as their observed substitutes, the $\hat{z}_k$-s. We let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})^T \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, 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 $z \in E = \{1, \ldots, K\}$, and

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

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

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

The proof of Theorem 7 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:

1. 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.

2. 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).

3. 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.

4. 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 8 (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*

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

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

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

Note that Assumption 4(2) implies that $\text{sep}(p) \to \infty$ 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 9 (Strong separation) *We say that the mixture model satisfies* strong separation *if there exists an $\varepsilon > 0$ such that $\text{sep}(p) \geq \varepsilon p$ eventually.*

Strong separation is equivalent to

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

A sufficient condition for strong separation is that for some $\varepsilon > 0$, $|\mu_i(z) - \mu_i(v)| \geq \varepsilon$ eventually for all $z, v \in E$, $v \neq z$. 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

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

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 $\text{sep}(p)$ 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 10 *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 $R_{z,v}^{(p)} \leq \frac{1}{10}$ for all $z, v \in E$ with $v \neq z$. Then

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

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

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

Remark 7 The proof of Proposition 10 is in Appendix A.3. It shows that the specific constants, 25 and 50, appearing in the bounds above hinge on the specific bound, $R_{z,v}^{(p)} \leq \frac{1}{10}$, 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.

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.

The upshot of Proposition 10 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) decreases 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

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

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 data set $\mathcal{S}_0$ in Algorithm 3 consists of i.i.d. observations of $\mathbf{X}_{1:p}$, the data set $\mathcal{S}$ 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 11 *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), $\widehat{\beta}_i^{\text{sub}} \overset{as}{\sim} \mathcal{N}(\beta_i, w_i^2/n)$, where the asymptotic variance parameter w_i^2 is the same as for the oracle estimator $\widehat{\beta}_i$.

Remark 8 The proof of Theorem 11 is in Appendix A.4. As mentioned in Remark 7, 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 9 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 $m, p \to \infty$, which is 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 by Anandkumar et al. (2014) or the derivations by Guo et al. (2022a) on page 2.

Guo et al. (2022a) 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 by Guo et al. (2022a) 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 by 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 proposed by 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.

1. We set $K = 10$ and fix $p_{\max} = 1000$ and $n_{\max} = 1000$.

2. We draw $\mu_i(z)$-s independently and uniformly from $(-1, 1)$ for $z \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, p_{\max}\}$.

3. 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.

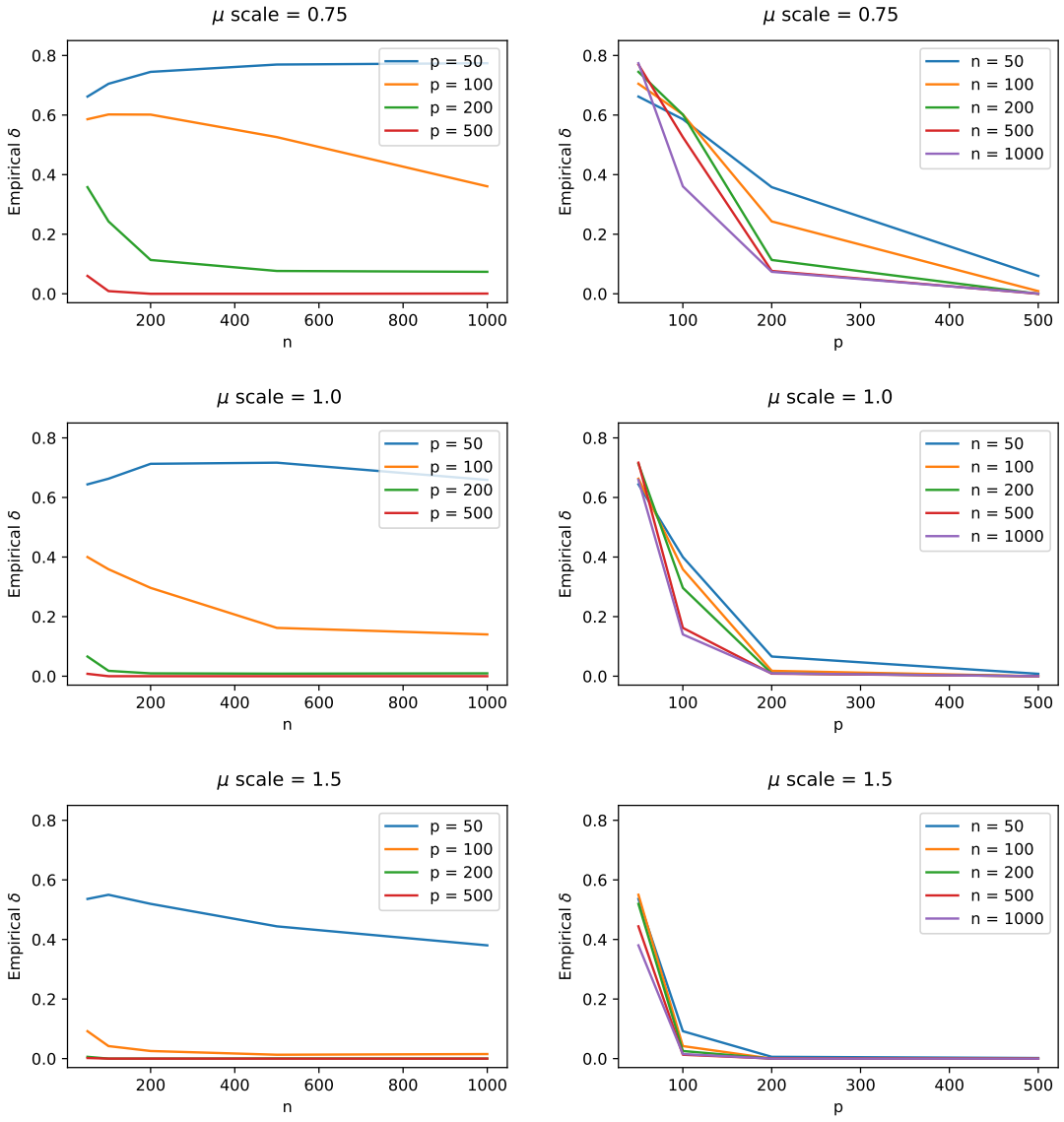

Figure 2: Empirical mislabeling rates as a function of $n = m$ and p and for three different separation scales.

We use the algorithm by 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 10 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.

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.

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

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

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

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

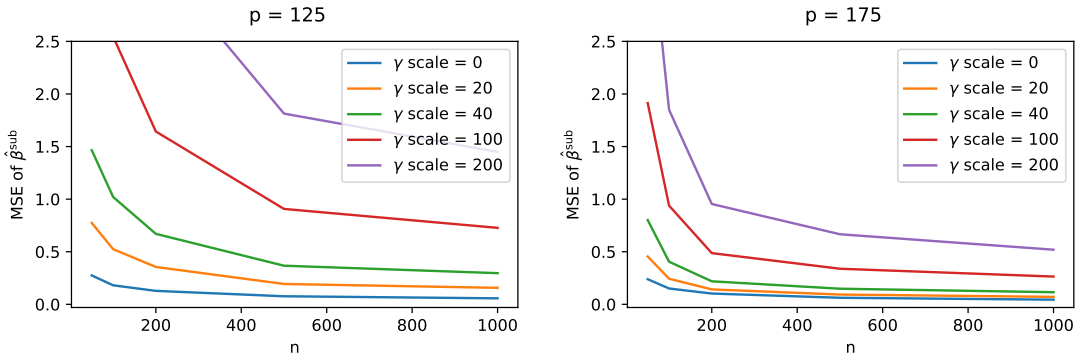

Figure 3: Average MSE for substitute adjustment using Algorithm 3 as a function of sample size n and for two different dimensions, a range of the unobserved confounding levels, and with $\mu_{\text{scale}} = 1$.

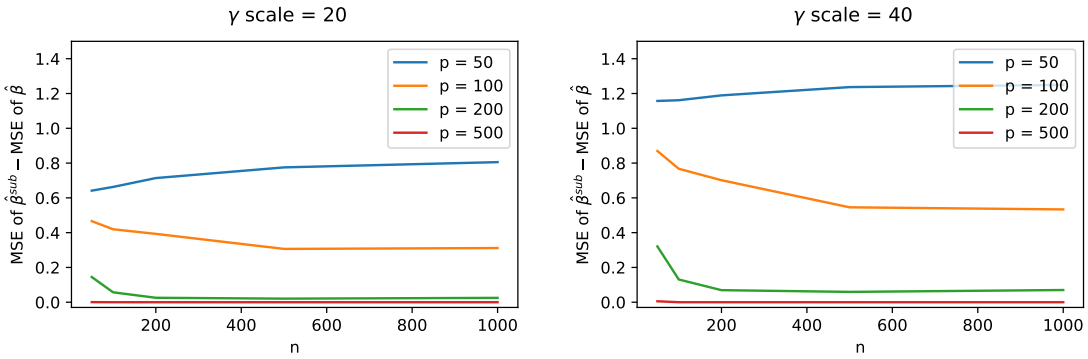

Figure 4: Difference in MSE between the substitute adjustement estimator, $\hat{\beta}^{\text{sub}}$, and the oracle estimator, $\hat{\beta}$, for $\mu_{\text{scale}} = 1$.

Figure 3 shows the MSE for the different combinations of n and p and for different choices of γ_{scale}. Unsurprisingly, the MSE decreases 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.

Our main theoretical result, Theorem 11, gives conditions under which the substitute adjustment estimator is asymptotically equivalent to the oracle estimator. Notably, the conditions ensure that the mislabeling rate tends to 0 sufficiently fast. Since we know the Z-s in the simulation study we can compute the oracle estimator and its average MSE. Figure 4 shows the difference in MSE by using substitutes for $\mu_{\text{scale}} = 1$. Unsurprisingly, the MSE of the substitute adjustment estimator is largest—but the difference almost vanishes when p is 200 or larger and the sample size is large enough. This aligns with the mislabeling rates also being small in this case, see Figure 2.

4.3 Comparisons with alternative estimators

Algorithm 3 implements substitute adjustment in its most obvious way; by plugging in the recovered values of the Z-s in the OLS estimator. In this section we compare Algorithm 3 with five other estimators. Three of these carry out the adjustment by regression on all the observed X_i-s, and two estimators augment these regression adjustment models by including the substitutes as regression variables.

1. *Ridge and focal ridge regression.* 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

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

   with λ chosen by five-fold cross-validation. The *focal ridge* regression estimator is obtained by considering separate ridge regression estimators for each focal parameter β_i—leaving out β_i from the penalization. That is,

$$\hat{\beta}^{(n,p)}_{i,\text{Focal-Ridge}} = \underset{\beta_i\in\mathbb{R}}{\arg\min} \min_{\substack{\beta_0\in\mathbb{R} \\ \boldsymbol{\beta}_{-i}\in\mathbb{R}^{p-1}}} \|\mathbf{y} - \beta_0 - \mathbb{X}_{-i}\boldsymbol{\beta}_{-i} - \mathbb{X}_i\beta_i\|_2^2 + \lambda\left\|\boldsymbol{\beta}_{-i}\right\|_2^2.$$

   To avoid excessive computations, the penalty parameter for focal ridge is chosen as $\lambda^{(n,p)}_{\text{Ridge}}$ for all i, where $\lambda^{(n,p)}_{\text{Ridge}}$ is the penalty parameter found by cross-validation for ridge regression.

2. *Augmented and focal augmented ridge regression.* Letting $\hat{\mathbf{Z}}$ denote the $n \times K$ model matrix of dummy variable encodings of the substitutes, the *augmented ridge* regression estimator is given as

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

   Note that the $\boldsymbol{\gamma}$-parameter is not penalized. Again, λ is chosen by five-fold cross-validation. The *focal augmented* ridge regression estimator is defined as above by leaving out β_i from the penalization when β_i is the focal parameter, and with the penalty parameter fixed as $\lambda^{(n,p)}_{\text{Aug-Ridge}}$ for all i.

3. *Double ML.* A double machine learning estimator is obtained by replacing the estimates $\hat{\mu}_i$ and $\hat{g}$, that are functions of the substitutes in Algorithm 3, by estimates of the regression functions

$$\mu_i^{\text{DML}}(\mathbf{x}_{-i}) = \mathbb{E}[X_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}]$$
$$g_i^{\text{DML}}(\mathbf{x}_{-i}) = \mathbb{E}[Y \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}].$$

   We report results with these nuisance functions estimated using gradient-boosted regression trees (GBTs). To avoid excessive computations, hyperparameters for learning all p GBTs g_i^{DML} were fixed and selected by a single five-fold cross-validation for learning g_1^{DML}. Similarly, hyperparameters for learning all p GBTs μ_i^{DML} were fixed and selected by five-fold cross-validation for learning μ_1^{DML}.

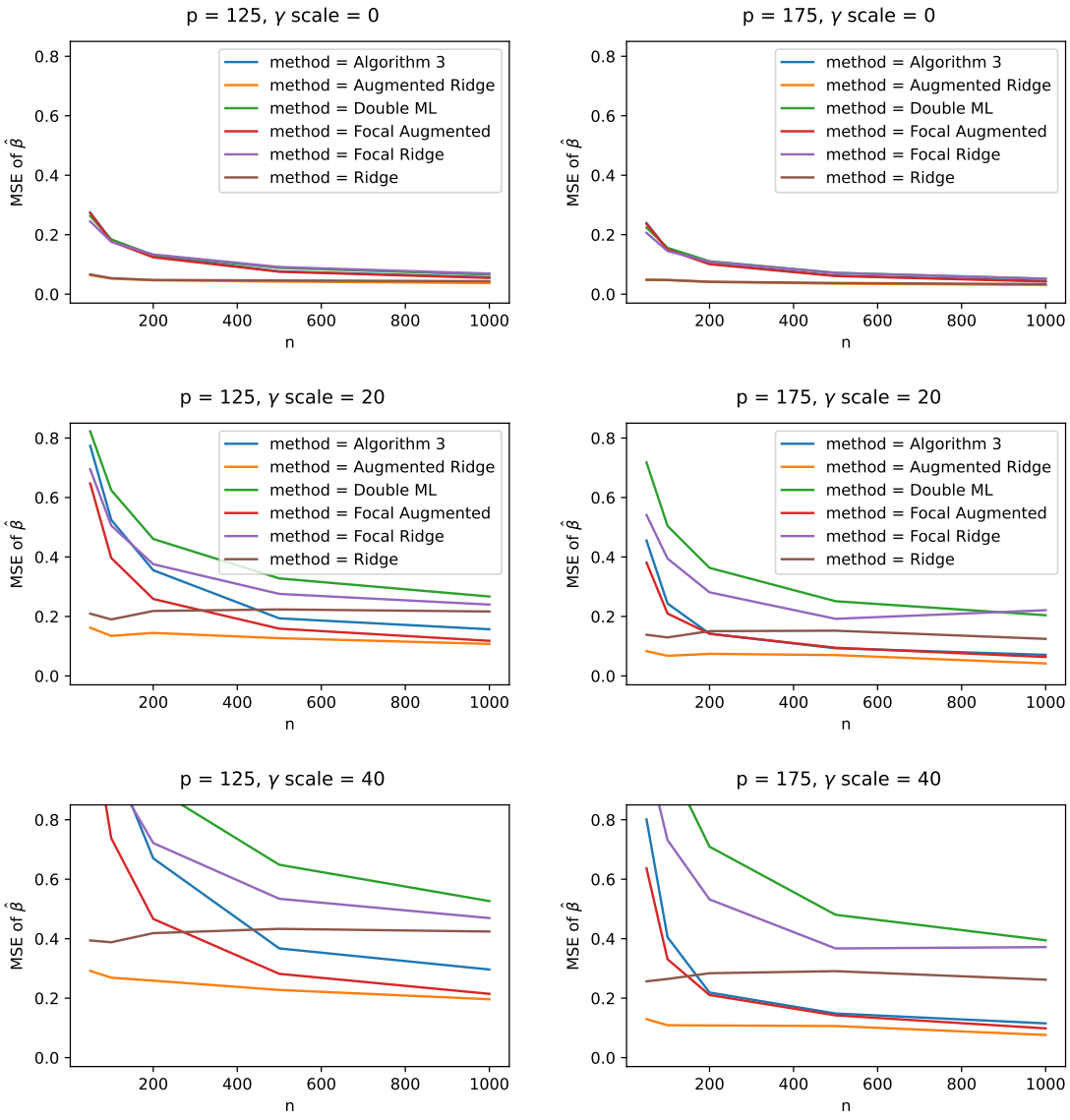

Figure 5: Average MSE for substitute adjustment using Algorithm 3 and five alternative estimators. Results are shown for $\mu_{\text{scale}} = 1$, two different dimensions, and three different levels of unobserved confounding.

The average MSE is computed for all five alternative estimators just as for substitute adjustment. Figure 5 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. The most important observations are summarized as follows:

1. The ridge and augmented ridge regression estimators have fairly small MSEs, which change little with sample size. The MSE is dominated by the bias induced by penalization, which leads to the favorable MSE, in particular for small sample sizes. This is most pronounced for the augmented estimator when there is unobserved confounding ($\gamma_{\text{scale}} > 0$),

2. For the remaining four estimators the MSE decreases as expected with sample size with almost no differences between the four estimators for $\gamma_{\text{scale}} = 0$ but notable differences for $\gamma_{\text{scale}} > 0$.

3. Among the four estimators, the focal augmented ridge regression estimator has the smallest MSE in all cases when $\gamma_{\text{scale}} > 0$. Algorithm 3 has a similarly small MSE when $p = 175$ (where the mislabeling rate is small), but suffers a little from the higher mislabeling rate for $p = 125$.

4. The focal ridge regression estimator and the double machine learning estimator have the largest MSE when $\gamma_{\text{scale}} > 0$. They appear to become increasingly worse with an increasing amount of unobserved confounding.

The good performance of the ridge and, in particular, the augmented ridge estimator should not be overinterpreted. Their biases are favorable to our particular simulation setup, as the relatively poor performance of the focal ridge estimator suggests, and the results are not likely to generalize. The three best-performing estimators all leverage the substitutes, with the augmented and focal augmented estimators having better small sample performance than Algorithm 3. Note that it is unsurprising that Algorithm 3 performs similar to the augmented and focal augmented estimators for large sample sizes and large p, 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.

4.4 Concluding remarks concerning the simulation study

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 10 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. Finally, we implemented double machine learning using random forests for estimating the nuisance regression functions, but the results were significantly worse than using GBTs.

The purpose of this simulation study is to support the asymptotic theory by investigating the finite-sample performance of substitute adjustment and comparing it to its direct

competitors. Obviously, we could break the performance of Algorithm 3 and the augmented ridge estimators by violating the model setup, e.g., by a more complicated structure of the latent variable, by additional dependence structure among the X_i-s, or by a more complicated outcome regression model. Although it is of interest to investigate this breakdown and, more importantly, how to alleviate a resulting performance loss, this is beyond the scope of the present paper.

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 comparable or better than alternative methods based on, e.g., naive linear or ridge regression as well as certain implementations of double machine learning. The better performance is achieved in settings where the unobserved confounding is sufficiently large and recovery is sufficiently good.

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_x^i 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 $\chi_x^i - \chi_{x_0}^i$ has a causal interpretation as an average treatment effect. The framework by Wang and 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 and 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 by Wang and 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 by Wang and 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 by Wang and 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 this depends crucially on the context and the causal assumptions we are willing to make. The factor model might be unrealistic or it might be implausible that we can recover the latent variable, for instance if p is small. In such cases the use of proxy or auxiliary variable methods, as considered by Louizos et al. (2017); Miao et al. (2018, 2023); Tchetgen et al. (2024), is likely more appropriate.

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. For once, the recovery Assumption 2(2) is a more plausible modeling assumption than recovery from a finite number of variables, and 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" as Wang and Blei (2020) require. 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.

Additionally, our infinite variable model gives a clear qualitative distinction between the adjusted mean of one (or any finite number of) variables and regression on all variables. According to Wang and Blei (2019), the deconfounder algorithm not only recovers the latent variable from all causes but it also estimates the effect of any *joint* intervention on all causes. In our view, this is too ambitious and leads to the counterexamples by D'Amour (2019). Our substitute adjustment algorithms only target one variable at a time in the adjusted regression. The joint distribution of (X_i, Z) still needs to be non-degenerate, though, to avoid the Section 6.5 counterexample by D'Amour (2019). Assumption 2(2) is not enough as it does not rule out exact recovery of Z from a single X_i. For the assumption-lean parameter, the positivity condition $\mathbb{E}[\mathrm{Var}[X_i \mid Z]] > 0$ in Assumption 3 serves this purpose. We emphasize that the case of primary interest is when Z can only be recovered exactly from an infinite number of variables, which makes this assumption benign.

We argue that substitute adjustment (and the deconfounder) should be used to target the adjusted mean χ^i_x, where you adjust for all other variables except X_i. Grimmer et al. (2023) come to a similar conclusion and 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. For $i = 1$, 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 with 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. Our simulation study actually illustrates this point. Since the mislabeling rates (see Figure 2) decrease with sample size, the recovery map $f^{(p)}$ would be estimated more accurately with access to unlabeled data. 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 and some practical advice

We believe that our error bound in Theorem 7 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. Kallus et al. (2018) present a related analysis of OLS estimation errors due to adjustment by regressors with errors.

We see two possible extensions of our result. For one, Theorem 7 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. It would be interesting to clarify how this relates to the literature on using proxy variables.

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. Heinrich and Kahn (2018) give more refined results on finite mixture estimation, and Ndaoud (2022) derives an optimal recovery method when $K = 2$ and the mixture distributions are Gaussian. In cases where the mixture distributions are Gaussian, it is also plausible that specialized algorithms as those by Kalai et al. (2012) and Gandhi and 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.

We believe that our theoretical results and simulations show that substitute adjustment can be a viable method for adjusted regression—but we acknowledge that the distributional

assumptions are strong and can be difficult to justify in practice. The purpose of this paper is not to advocate uncritical usage of substitute adjustment but to clarify when it actually works. Besides a correctly specified latent variable model it is also important that the recovery error is sufficiently small, which would typically require p to be large. This is also when substitute adjustment has the most obvious computational and statistical benefit over naive regression, say. If p is small, adjustment via a semiparametric regression model is likely a better choice, or we might want to consider proxy variable methods to additionally justify a causal interpretation.

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 2 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 $\tilde{B} = f_i^{-1}(B) \subseteq \mathbb{R}^{\mathbb{N}}$,

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

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

We finally find that

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

■

Proof of Proposition 5 We find that

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

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 $\operatorname{Cov}\left[X_i, b^i(Z) \mid Z\right] = 0$, whence $\beta_i = 0$.

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

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

and (10) follows. ■

A.2 Auxiliary results related to Section 3.2 and proof of Theorem 7

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 7 is based on the following bound on the difference between the projection matrices.

Lemma 12 *Let α and δ be as defined by* (16) *and* (17)*. If $\alpha > 0$ it holds that*

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

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),

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

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

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

We also have that

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

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 7, note that

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

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

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

Proof of Theorem 7 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 $\rho \leq 1$,

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

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

A.3 Auxiliary concentration inequalities. Proofs of Propositions 6 and 10

Lemma 13 *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 $R_{z,v}^{(p)} \leq \frac{1}{10}$ for all $z, v \in E$ with $v \neq z$ then*

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

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

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

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

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

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

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

This gives

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

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

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 $\mathbb{E}[W_i \mid Z=z] = 0$ 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

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

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

Before proceeding to the concentration inequality for sub-Gaussian distributions, we use Lemma 13 to prove Proposition 6.

Proof of Proposition 6 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 13 gives

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

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 $\mathbb{P}\left(\hat{Z}_{p_r} \neq Z\right) \leq \frac{1}{r^2}$. 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

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

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 14 *Consider the same setup as in Lemma 13, that is, Assumption 4 holds and $R_{z,v}^{(p)} \leq \frac{1}{10}$ 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*

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

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

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

for $\lambda \in \mathbb{R}$. Consequently, with W_i as in the proof of Lemma 13, and using conditional independence of the X_i-s given $Z = z$,

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

Using (31) in combination with the Chernoff bound gives

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

where we, as in the proof of Lemma 13, have used that the bound on the relative error implies that $B_{z,v} \leq \frac{6}{5}$. ■

Proof of Proposition 10 The argument proceeds as in the proof of Proposition 6. 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 13 then gives

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

If the sub-Gaussian assumption holds, Lemma 14 instead gives

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

■

A.4 Proof of Theorem 11

Proof of Theorem 11 Recall that

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

hence by Proposition 10

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

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

Now rewrite the bound (19) as

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

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

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

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 $z \in E$

$$|\hat{n}(z) - n(z)| \leq \sum_{k=1}^{n} |1(\hat{z}_k = z) - 1(z_k = z)| \leq \sum_{k=1}^{n} 1(\hat{z}_k \neq z_k) \leq n\delta.$$

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 12,

$$\begin{aligned}\left|\|(I - P_{\hat{\mathbf{Z}}})\mathbf{x}_i\|_2^2 - \|(I - P_{\mathbf{Z}})\mathbf{x}_i\|_2^2\right| &= \left|\|(P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}})\mathbf{x}_i\|_2^2 + 2|\langle (I - P_{\hat{\mathbf{Z}}})\mathbf{x}_i, (P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}})\mathbf{x}_i\rangle\right| \\ &\leq \|P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}}\|_2^2\|\mathbf{x}_i\|_2^2 + 2\|P_{\hat{\mathbf{Z}}} - P_{\mathbf{Z}}\|_2\|\mathbf{x}_i\|_2^2 \\ &\leq \left(\frac{2\delta}{\alpha} + \sqrt{\frac{2\delta}{\alpha}}\right)\|\mathbf{x}_i\|_2^2.\end{aligned}$$

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 $\frac{\text{sep}(p)}{n} \to \infty$. Then

$$\sqrt{n}|\widehat{\beta}_i^{\text{sub}} - \hat{\beta}_i| \leq \sqrt{n\delta}L_n$$

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 $h_n = \frac{\text{sep}(p)}{\log(n)} \to \infty$, then we can replace (33) by the bound

$$\mathbb{E}[\delta] \leq K\exp\left(-\frac{\text{sep}(p)}{50v_{\max}}\right) + K^2\max_{z \neq v}\mathbb{P}\left(R_{z,v}^{(p)} > \tfrac{1}{10}\right).$$

Multiplying by n, we get that the first term in the bound equals

$$\begin{aligned}Kn\exp\left(-\frac{\text{sep}(p)}{50v_{\max}}\right) &= K\exp\left(-\frac{\text{sep}(p)}{50v_{\max}} + \log(n)\right) \\ &= K\exp\left(\log(n)\left(1 - \frac{h_n}{50v_{\max}}\right)\right) \to 0\end{aligned}$$

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 15 (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 $-\log\bigl(\mathrm{BC}_{z,v}^i\bigr)$ 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 16 *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 15. ■

Corollary 17 *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 $A_z \cap A_v = \emptyset$ 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 $v \neq z$

$$\begin{aligned} P_v(A_z) &= \frac{\mathbb{P}(\mathbf{X} \in A_z, Z = v)}{\mathbb{P}(Z = v)} = \frac{\mathbb{P}(f(\mathbf{X}) = z, Z = v)}{\mathbb{P}(Z = v)} \\ &= \frac{\mathbb{P}(f(\mathbf{X}) \neq Z, Z = v)}{\mathbb{P}(Z = v)} = \frac{0}{\mathbb{P}(Z = v)} = 0. \end{aligned}$$

This shows that P_z and P_v are singular, and by Proposition 16, either (36) or (37) holds. ■

References

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Richard Berk, Andreas Buja, Lawrence Brown, Edward George, Arun Kumar Kuchibhotla, Weijie Su, and Linda Zhao. Assumption lean regression. *The American Statistician*, 75 (1):76–84, 2021.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020.

William G Cochran. Errors of measurement in statistics. *Technometrics*, 10(4):637–666, 1968.

James Durbin. Errors in variables. *Review of the International Statistical Institute*, 22(1/3): 23–32, 1954.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3478–3486, 2019.

Kavish Gandhi and Yonah Borns-Weil. Moment-based learning of mixture distributions. *MIT Summer Program for Undergraduate Research (SPUR)*, 2016.

Justin Grimmer, Dean Knox, and Brandon 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.

Bingni Guo, Jiawang Nie, and Zi Yang. Learning diagonal Gaussian mixture models and incomplete tensor decompositions. *Vietnam Journal of Mathematics*, 50:421–446, 2022a.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, 50(3):1320 – 1347, 2022b.

Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844 – 2870, 2018.

Shizuo Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, 49 (1):214–224, 1948.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

Olav Kallenberg. *Foundations of Modern Probability*. Probability and Stochastic Modelling. Springer-Verlag, New York, third edition, 2021.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal Inference with Noisy and Missing Covariates via Matrix Factorization. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, 3(9):1–12, 2007.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

Anton Rask Lundborg, Ilmun Kim, Rajen D Shah, and Richard J Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851 – 2878, 2024.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Wang Miao, Wenjie Hu, Elizabeth L. Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023.

Mohamed Ndaoud. Sharp optimal recovery in the two component Gaussian mixture model. *The Annals of Statistics*, 50(4):2096 – 2126, 2022.

Elizabeth L Ogburn, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Counterexamples to "The blessings of multiple causes" by Wang and Blei. *arXiv:2001.06555*, 2020.

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLOS Genetics*, 2(12):1–20, 2006.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Susanne M Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377, 2016.

Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554, 2015.

Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Stijn Vansteelandt and Oliver Dukes. Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):657–685, 2022.

Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B Owen. Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics*, 45(5):1863–1894, 2017.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Yixin Wang and David M Blei. Towards clarifying the theory of the deconfounder. *arXiv:2003.04948*, 2020.