Unmeasured Confounding in High-Dimensional Data

Overview

Unmeasured confounding is a strong limitation to drawing causal inferences and, in general, a key threat to observational studies. This essay explores multiple novel approaches to overcome this problem and focuses on the case of high-dimensional data. Following the recent and rich literature on the subject, it presents two main methods based on linear confounding models. The first looks at the task of high-dimensional linear regression and studies how adjusting the singular values of the observed design matrix can remove the confounding effect from the data under some assumptions. The second deals with multiple hypothesis testing and relies on linear factor models to obtain estimators of the underlying effect and valid asymptotic tests despite unmeasured confounding. A brief introduction to recent developments using non-linear models and the prolific discussions they sparked in this research area is also presented.

This essay relies for its vast majority on the articles that are cited and presented along the way. The personal contributions are located for the most part in the third and fourth sections. They consist principally in an investigation of the alignment between the confounding bias and the first right singular vector of the design matrix in a linear confounding model with one confounder; in empirical simulations that illustrate the role of this alignment in the efficacy of the confounding adjustment method; and in two small case studies on how to adjust for confounding in practice using the methods presented, applied to a real dataset on the determinants and the impacts of tourism in German cities.

All empirical simulations, along with data for the case study, can be found in a GitHub repository (`https://github.com/bglbrt/UCHDD/`) whose architecture is detailed in the appendix of this essay.

Table of contents

1 Introduction

1.1 Unmeasured confounding

We briefly introduce the concept of confounding with an example. Suppose that a mayor is interested in developing tourism in her municipality. To do so, she collected data on many nearby towns, including variables regarding demographics, economic activity, etc. In her research, she observes that the number of tourists per year is highly correlated with the number of fishers per town. She reflects that it is unlikely that tourists are particularly interested in the craft of fishers, nor that these usually prefer to live in touristic places. She also thinks it would be rather unlikely that increasing the number of fishers working in her city would lead to an increase in tourists. Comparably, it is much more sensible that seaside access in coastal cities, which she did not record, drives both tourism and fishing activity.

This example illustrates the concept of confounding. In this case, the effect of the number of fishers in a city on its touristic appeal is confounded by the city's access to the seashore, which is most likely the primary reason for the tourists visiting. In general terms, confounding is a causal concept that can be described as such: when interested in the causal effect of a treatment on an outcome, confounding is defined as the bias caused by the shared causes of the treatment and the outcome. When these shared causes are not accessible to the researcher, confounding is said to be unmeasured. The simplest scenario of unmeasured confounding can be illustrated in Figure 1 using a directed acyclic graph (DAG), where one is interested in the causal effect of an explanatory variable X on an outcome variable Y but where the variable Z is an unmeasured confounder between X and Y. Here, notice that the fact that Z is unmeasured is visually represented by a dashed circle and dashed arrows.

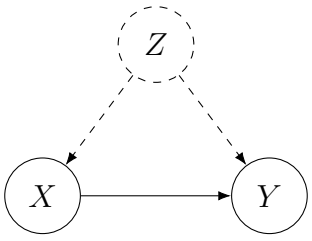


Figure 1: Unmeasured confounding between X and Y

In a broader sense, one may not necessarily wish, or be able to place oneself in a causal framework, and may only be interested in exploring the associations between some covariates and an outcome (for instance, using regression). In this setting, with a slight abuse of terminology, unmeasured confounding can be seen as the case where some unobserved variables correlate with both the explanatory variables and the outcome variables. When not accounted for, these unobserved covariates can lead to misleading conclusions and biased estimates. These misleading conclusions are often referred to as spurious correlations or associations, and the unobserved covariates as confounders, extraneous factors, or lurking variables.

Hence, when interested in causal inference, a key assumption, along with the usual ones of *consistency*, *no interference* and *positivity*, is that of *no unmeasured confounders*. In fact, when one can find a set of covariates that satisfies the *no unmeasured confounders* assumption, one can show that the causal effect of the treatment on the outcome can be identified. However, this setting requires the strong assumption that all possible associations between covariates can be explained from the observed set of covariates. Having such a set of covariates at hand can be difficult in practice and inherently depends on the investigator's beliefs on the underlying causal relationships. Furthermore, even when the researcher is confident that all sources of confounding have been observed and are accounted for, there is no way to assess with certainty that the *no unmeasured confounders* assumption holds.

When the objective doesn't necessarily lie in finding the true causal relationships between covariates, confounding can be just as much of an issue, and results are often commented under the implicit assumption that no unobserved covariate is responsible for the studied associations. Thus, unmeasured confounding is one of the major concerns when studying the associations, or causal relationships, between variables.

1.2 Dealing with confounding in practice

In practice, the fact that all potential confounders are observed can be argued to be more or less likely to hold given the context and the data. However, because it is inherently unverifiable, multiple statistical approaches such as sensitivity analysis and instrumental variables have been put forward both to assess the robustness of inferences under potential confounding and to offer guarantees for identification of the underlying effect under additional assumptions.

When studying the relationship between two covariates, instrumental variables are widely used in social and medical sciences, both in models that allow inferring causal relationships or in less restrictive estimation procedures. However, when interested in the associations between a large number of covariates and outcomes, although instrumental variables can be applied theoretically, using them may be much less practical. In fact, to keep the identification properties (for instance, the identification of the causal effect of interest under the right assumptions), the researcher needs to find an instrumental variable for each covariate that is potentially confounded. Yet, modern datasets like gene expression datasets contain thousands of covariates, for which finding adequate instrumental variables may not be feasible. Building on this and other limitations, multiple approaches are currently being researched to alleviate some of the statistical issues that come with unmeasured confounding.

In this essay, we focus on how high-dimensional statistics and datasets present new challenges – and opportunities – for dealing with unmeasured confounding. To do so, we try to review a few recent papers and the novel statistical approaches they introduce that aim at leveraging aspects of high-dimensionality for estimation under unmeasured confounding and explore the theoretical guarantees they offer. We see that the underlying (unconfounded) effect of interest is identifiable and can be estimated from the data under specific assumptions.

1.3 Leveraging the structure of high-dimensional data

High-dimensional statistics refer to the setting where the number of unknown parameters to estimate is larger than the number of samples. This broad definition encompasses many statistical settings like supervised regression with more covariates than samples or multiple hypothesis testing where the number of tests exceeds the number of samples. In this introductory section, we try to give some very first intuition on the specificity of high-dimensional data under potential unmeasured confounding.

Suppose first that the task being considered is that of estimating the effect of a large number of explanatory variables on a single outcome variable, for instance, using a regression model. Suppose also that the researcher expects that the potential confounders affect a lot or all of the explanatory variables. Then, because the potential confounders affect many of the covariates, one can expect that some information about the confounders can be retrieved from these observed covariates. In other words, because the confounder induces variations in each measured explanatory variable, one can hope that the confounder can be implicitly learned from the observed data. The aim would then be to use this information on the confounders to adjust for their effect. An instance of this scenario is illustrated using a DAG in Figure 2, where one is interested in the causal effect of $X_1, X_2, ..., X_p$ on the outcome variable Y but where for all $i \in \{1, ..., p\}$, Z is an unmeasured confounder between X_i and Y.

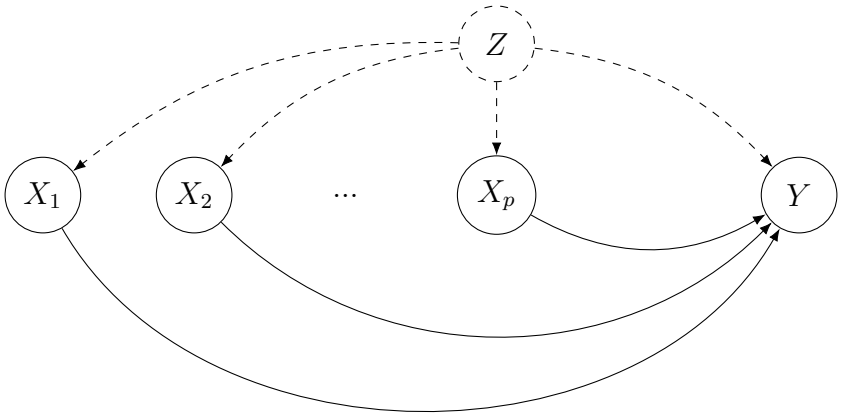


Figure 2: Unmeasured confounding with many covariates and one outcome variable

The same reasoning can be applied when the task considered is that of testing the effect of a single covariate on a large number of outcome variables, for instance, in the context of multiple hypothesis testing. In this setting, suppose that one expects all the potential confounders to affect a lot or all of the outcome variables. Then, again, because the potential confounders affect many of the observed variables, one can imagine that the confounder can be implicitly learned from the observed outcome variables and then adjusted for in the estimating procedure. We illustrate this case in Figure 3 (on page 4), where one is interested in the causal effect of the observed covariate X on the outcome variables $Y_1, Y_2, ..., Y_q$ but where for all $i \in \{1, ..., q\}$, Z is an unmeasured confounder between X and Y_i.

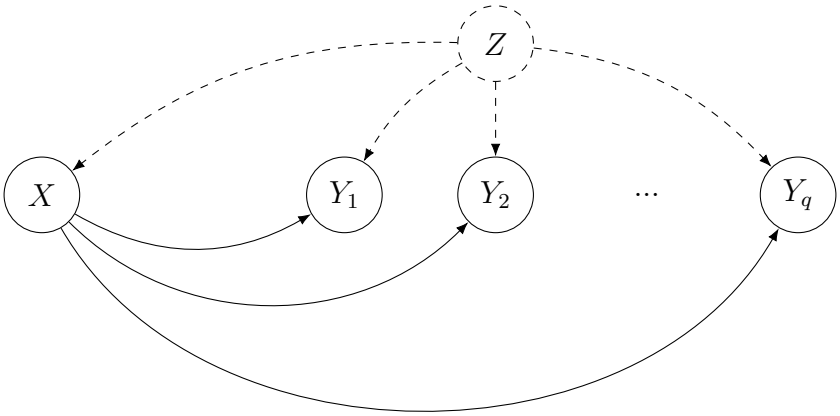


Figure 3: Unmeasured confounding with many outcome variables and one covariate

In essence, the approaches presented in this essay try to build on this intuition to study when and how some information about the confounders can be recovered from the observed covariates or outcomes.

In short, these approaches can be grouped under the idea of using a two-step estimation procedure to adjust for confounding. This procedure can be described in a very general way as follows. In the first step, some quantity about the confounding variables is estimated from the observed data (for instance, some latent variables). Then, in the second step, the observed data is adjusted for this quantity. How the estimated quantity that is adjusted for in the second step is defined varies between approaches. In this essay, we focus primarily on the two following strategies.

The first comes from the idea and heuristics that much of the information about the confounders can be found in the first few principal components of the observed data and can be seen as a generalisation of this idea. Thus, in the procedure described above, the quantity estimated in the first step can be regarded here as the first few principal components of the observed data, which can then be adjusted for in different ways. In the terms of Ćevid et al. (2018), this approach can be seen as implicit because it does not attempt to directly estimate surrogate variables for the confounders and to use them as additional covariates in the model.

The second uses exploratory factor models and thus posits latent variables called factors, which are estimated from the data and which we can think of as surrogate variables for the unobserved confounders. Notice, however, that one may argue that this distinction is slightly artificial, as, broadly speaking, both methods attempt to provide a low-rank approximation of a given covariance matrix and to use this representation of the data to estimate and adjust for potential confounding.

To illustrate where these results may be applied practically, we introduce two concrete examples of potential unmeasured confounding in high-dimensional data, as well as some rough intuition as to why, in these real-life situations, high-dimensional data may offer new opportunities for dealing with unmeasured confounding.

Example E1 (Genetic determinants of cardiovascular disease)**.** Suppose that a researcher is interested in finding how changes in an individual's genome affect his probability of developing cardiovascular diseases. To do so, she has access to an SNP Genotyping dataset. The dataset encodes which variations of the genome were observed for each individual whose genome was sequenced in the study. To identify the genetic variations associated with an increase in the probability of developing cardiovascular disease, the researcher is interested in doing multiple hypothesis testing. However, ethnicity is known in epidemiological studies to affect both the genome of the individuals as well as the probability of developing cardiovascular disease (because it incorporates many other biological and cultural dimensions). Thus, ethnicity can be seen as a potential confounder in this study.

In this setting, if ethnicity or some proxy of it was not measured as part of the study, the researcher is presented with a potential case of unmeasured confounding. However, following the intuition given above, she may hope to infer some information about the confounder from the large number of genetic variations measured in the study, or, in other words, that these genetic variations contain some information on the individual's geographical origin, which she may then want to exploit.

Example E2 (Determinants of tourism in a country)**.** Following our introductory example, suppose that a researcher is interested in what makes a city attract tourists within a country. To do so, she has access to a dataset measuring many characteristics of the studied cities, such that there are more variables than observations. However, as we argued, the relationship is likely to be confounded by unobserved (and non-measurable) characteristics. Hence, in the same way, the researcher is presented with a likely case of unmeasured confounding and may expect to retrieve some information on the confounders from the observed data.

1.4 Contents and outline

In the following sections, we try to present multiple and different approaches to this problem, and when possible, to illustrate these approaches with examples, simulations, or by applying the estimating procedures to a real dataset. We always introduce these approaches along with references to the articles in which they were derived.

In Section 2, we start by introducing the notations that will be used throughout this essay and by presenting a general framework for modelling confounding, both using counterfactual notations and in a non-necessarily structural linear setting. In Section 3, we focus on the case of regression of multiple covariates on a single outcome. To do this, we first present the approach developed by Ćevid et al. (2018) and Guo et al. (2020) on spectral transformations, which generalise the idea of adjusting for the first few principal components of the observed data. We also include a much shorter discussion on recent developments in non-linear approaches using probabilistic factor models, based on the work of Wang and Blei (2019) and the rich literature commenting on their findings. Finally, in Section 4, we focus on multiple hypothesis testing and present the work of Wang et al. (2017), who group multiple adjustment approaches under the same framework and provide theoretical guarantees for identifying the underlying effect and control the overall Type I error under some assumptions.

2 Preliminaries and framework

2.1 Notations and definitions

We start by introducing some notations that will be used throughout this essay. We usually write n for the number of samples. For a generic observation in the sample, we denote by $X = (X_1, ..., X_p)$ a vector of p covariates which are our explanatory variables or primary variables of interest; and by $Y = (Y_1, ..., Y_q)$ a vector of q outcome or response variables. In general, when interested in the task of learning how multiple covariates affect a single outcome, this reduces to $q = 1$, and we will write the single outcome variable as Y. When interested in doing multiple hypothesis testing to learn about the effect of a single covariate on multiple outcomes, this reduces to $p = 1$, and we will write the single covariate as X. Then, we denote by $Z = (Z_1, ..., Z_r)$ a vector of r unobserved covariates which correlate with both the primary variables of interest and the outcome variables and which we refer to as confounders. Additionally, we denote by $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ the supports of X, Y and Z respectively.

Then, for any random vectors $A \in \mathbb{R}^p$ and $B \in \mathbb{R}^q$, we write $\mathrm{Cov}(A) \in \mathbb{R}^{p\times p}$ for the covariance matrix of A, and $\mathrm{Cov}(A, B) \in \mathbb{R}^{p\times q}$ for the matrix whose entries $[\mathrm{Cov}(A, B)]_{i,j}$ are equal to the covariance between the i-th component of A and the j-th component of B. Finally, for generic A and B we use the following notations:

- $A = \mathcal{O}(B)$ if there exists a constant $c > 0$ such that, asymptotically, $A < cB$,
- $A = \mathcal{O}_p(B)$ if there exists a constant $c > 0$ such that $\mathbb{P}(A > cB) \to 0$,
- $A = \Omega(B)$ if there exists a constant $c > 0$ such that, asymptotically, $A > cB$,
- $A = \Omega_p(B)$ if there exists a constant $c > 0$ such that $\mathbb{P}(A \geq cB) \to 1$.

2.2 Modelling confounding

We begin by introducing more formally the challenges of identifying causal effects under unmeasured confounding using causal terminology, in a general non-parametric setting and using counterfactual notations. Then, we present the parametric and linear framework in which we will be most interested. In this last framework, we will focus on the statistical methods for estimating the effect of a covariate on an outcome without necessarily assuming structural or causal relationships between the variables.

2.2.1 A counterfactual approach to modelling confounding

Using causal inference terminology, one can think of the statistical tasks that we focus on as a problem where we are interested in learning the causal effect of the covariates X on the outcome Y, where this causal effect is confounded by some variables Z.

In this rather general setting, let $Y(x)$ be the counterfactual value of the outcome Y had the treatments X been set to x. Then, using f to denote a probability density or probability mass function, our main aim lies in identifying $f(Y(x))$ the distribution

of the outcome Y when the treatments X are set to a given value x; and in estimating quantities of interest regarding the potential causal relationships between X and Y, such as the treatment effect of a given covariate on an outcome variable.

When interested in studying how covariates are causally related to an observed outcome, we make the two following standard assumptions.

Assumption 1 (Consistency)**.** For a generic observation in the sample, when the covariates X are observed to have value x, the value of the counterfactual outcome $Y(x)$ is equal to that of the observed outcome Y:

$$\forall x \in \mathcal{X} : X = x \Rightarrow Y(x) = Y. \tag{A1}$$

Assumption 2 (Positivity)**.** For all potential values taken by the confounders z, for any value of the covariates x, any observation in the sample had a positive probability of having its observed covariates taking value x:

$$\forall (x, z) \in \mathcal{X} \times \mathcal{Z} : f(X = x \mid Z = z) > 0. \tag{A2}$$

Identification and estimation in this setting are difficult as the confounders Z affect both the outcome Y and the treatments X, so the interventional distribution $f(Y(x))$, which is the distribution of the potential outcome had X been set to x artificially by the researcher, differs from the conditional distribution $f(Y \mid X = x)$, which we can estimate from the observed data.

If all confounders were observed (that is, if Z was a sufficient subset of confounders and if Z was observed), then $f(Y(x))$ would be fully identifiable, as it could be written entirely in terms of observable distributions. Suppose for now that Z is observed and that we make the following additional assumption.

Assumption 3 (No unmeasured confounders)**.** Z is a sufficient set of covariates such that, conditional on these, X does not depend on the counterfactual outcomes $Y(x)$:

$$Y(x) \perp\!\!\!\perp X \mid Z. \tag{A3}$$

Under this additional assumption, if Z is observed, by using in the following order the law of total expectation, the *no unmeasured confounders* assumption, and the *consistency* assumption, we have $\forall y \in \mathcal{Y}$:

$$\begin{aligned} f(Y(x) = y) &= \mathbb{E}[f(Y(x) = y \mid Z)] \\ &= \mathbb{E}[f(Y = y \mid X = x, Z)] \\ &= \int_{\mathcal{Z}} f(Y = y \mid X = x, Z = z) f(Z = z)\, \mathrm{d}z. \end{aligned}$$

Hence, in the ideal setting where we observe Z and where the *no unmeasured confounders* assumption is verified, all quantities in the above equation can be estimated from the observed data, and $f(Y(x))$ is identifiable.

However, when the confounders are not observed, without further assumptions, one cannot identify $f(Y(x))$, because the joint distribution of Y, X and Z cannot be inferred from the observed data. In fact, using the notations from Miao et al. (2020), under the same assumptions but when one thinks of Z as unobserved, we can write:

$$f(Y(x) = y) = \int_{\mathcal{Z}} f(Y = y \mid X = x, Z = z) f(Z = z) \, \mathrm{d}z,$$

$$\text{where} \quad f(Z = z) = \int_{\mathcal{X}} f(Z = z \mid X = x) f(X = x) \, \mathrm{d}x = \int_{\mathcal{X}} f(Z = z, X = x) \, \mathrm{d}x.$$

Thus, we wish to determine both $f(Y = y \mid X = x, Z = z)$ and $f(X = x, Z = z)$, but, because Z is unobserved, only $f(X = x)$ and $f(Y = y \mid X = x)$ are known. For clarity, notice that we can write the following relationships:

$$f(X = x) = \int_{\mathcal{Z}} f(X = x, Z = z) \, \mathrm{d}z,$$

$$f(Y = y \mid X = x) = \int_{\mathcal{Z}} f(Y = y \mid X = x, Z = z) f(Z = z \mid X = x) \, \mathrm{d}z.$$

However, in general, the joint distribution cannot be uniquely determined using the marginal distribution, so $f(Y = y \mid X = x, Z = z)$ and $f(X = x, Z = z)$ cannot be identified. Thus, it makes intuitive sense that $f(Y(x))$ is not identifiable as such. And in fact, D'Amour (2019) showed using counter-examples that general non-parametric identification is impossible in this setting.

2.2.2 Linear statistical confounding models

In most of this essay, we reduce this framework to a parametric one where both the variables of interest X and the unmeasured confounders Z are assumed to have a linear effect on the outcomes Y, and where the variables of interest and the unmeasured confounders are also linearly related. We present such models in this section.

It is important to notice, however, that in this linear setting, one can only interpret β as the causal effect of the covariates X on the outcome Y if we assume that the relationships between the covariates are structural, i.e. if we interpret the models as structural equations models; and if the models are correctly specified. In general, we do not make structural assumptions on the relationships between the covariates, and thus the studied coefficients of interest are not regarded as causal effects.

Consider a linear model with n observations. In this essay, we only consider the two sub-cases where either $p = 1$ or $q = 1$. For clarity, we distinguish these two cases in models M1 and M2. When interested in the task of learning how multiple covariates affect a single outcome (so $q = 1$), for instance, using regression, we model the relationships between the covariates in the following way:

$$\begin{aligned} X_{n\times p} &= Z_{n\times r}\Gamma_{r\times p} + E_{n\times p}, \\ Y_{n\times 1} &= X_{n\times p}\beta_{p\times 1} + Z_{n\times r}\delta_{r\times 1} + \nu_{n\times 1}. \end{aligned} \tag{M1}$$

When interested in doing multiple hypothesis testing to learn about the effect of a single covariate on multiple outcomes (so $p = 1$), we will model them as such:

$$\begin{aligned} Z_{n\times r} &= X_{n\times 1}\alpha_{1\times r} + W_{n\times r}, \\ Y_{n\times q} &= X_{n\times 1}\beta_{1\times q} + Z_{n\times r}\Pi_{r\times q} + U_{n\times q}. \end{aligned} \tag{M2}$$

2.2.3 Exploratory factor models

An important component of some of the approaches we will focus on in this essay is the use of exploratory factor models. In this section, we present a very brief introduction to this statistical method and to how to fit it. In short, exploratory factor analysis aims to explain variability in the observed data using unobserved latent variables called factors, where the number of factors is generally lower than the number of observed covariates. As such, it can be seen and used as a data reduction technique, and, although the two are different, factor analysis is closely related to PCA. Formally, a linear factor analysis model can generally be written as:

$$X = LF + E.$$

In this model, $X \in \mathbb{R}^{n\times p}$ is an observed design matrix with centred columns, $L \in \mathbb{R}^{n\times r}$ is referred to as the factor loadings matrix, $F \in \mathbb{R}^{r\times p}$ as the factor matrix, and $E \in \mathbb{R}^{n\times p}$ as an error matrix. Notice that L, F and E are all unobserved. Here, r is the number of latent variables. For clarity, one can picture an example of a design matrix X consisting of the scores obtained by n restaurants in p different criteria, and where $r = 2$ with the two latent variables being the quality of food and the quality of service. However, in exploratory factor models, the r factors are only latent constructs that are posited, and cannot be interpreted as existing covariates.

Exploratory factor models are usually fitted either by principal axis factoring (PAF) or using maximum likelihood estimation (MLE). We illustrate how the latter can be done very briefly under distributional assumptions. Suppose that the rows of E are independent of the rows of F and distributed from a multivariate normal distribution $\mathcal{N}(0, \Psi)$ where Ψ is diagonal with entries $\psi_1, ..., \psi_p$. Suppose that all entries in F are distributed from a $\mathcal{N}(0, 1)$ distribution. Then, the rows of X follow a multivariate normal distribution $\mathcal{N}(0, \Psi + L^T L)$, and the factor loadings L and Ψ can be estimated using the following log-likelihood (for instance with the EM algorithm):

$$l(\Psi, L \mid X) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log\det(\Psi + L^T L) - \frac{1}{2}\sum_{i=1}^{n} X_i^T(\Psi + L^T L)^{-1}X_i.$$

Such models are closely related to the linear confounding models introduced above. Recall that, in model M1, Z is unobserved. Then, if X is column centred and if r is known, the first equation in model M1 is exactly that of an exploratory factor model. Thus, under some assumptions on the error matrix, an exploratory factor model can be used to estimate Z and Γ, where the columns of the estimated matrix for Z (or L in the model above) can be used as surrogate variables for the confounders.

3 Confounding in high-dimensional regression

We focus in this section on the task of measuring the effect of multiple covariates on a single outcome using high-dimensional regression. We present in detail the approach of Ćevid et al. (2018) based on spectral transformations and briefly introduce the non-linear approach taken by Wang and Blei (2019) along with some of the rich reviews and comments it sparked in this research area.

3.1 Adjusting for confounding using spectral transformations

In this first section, we derive a general framework for a set of methods using spectral transformations for point parameter estimation in a high-dimensional linear regression setting following Ćevid et al. (2018). Although we do not cover it in this essay, the follow-up work of Guo et al. (2020) allows one to go further and correct the biases induced by these spectral transformations and by high dimensionality, and on obtaining confidence intervals for the resulting estimates.

The method presented by Ćevid et al. (2018) builds on multiple approaches that have been proposed to alleviate some of the confounding-induced bias in observational studies and brings them together in the same framework. In broad terms, a spectral transformation is a linear transformation that modifies the spectrum of the matrix it is applied to. The most common examples are spectral transformations that affect the singular values. The procedure described by Ćevid et al. (2018) aims at reducing the bias due to unmeasured confounding by first applying some spectral transformation to the design matrix of the observed data, and then using a usual regression model for high-dimensional data like the Lasso on the transformed design matrix.

To gain some initial intuition on why one may use spectral transformations to adjust for unmeasured confounding, we start by presenting a concrete example where such methods are being used in practice.

Following on example E1, recall that we argued that ethnicity could potentially act as an unmeasured confounder in a study of the effect of the genetic determinants of cardiovascular diseases. Ideally, to adjust for this potential confounding, the researcher would want to have some information about the individuals' geographical origins in the recorded variables of the study. However, because ethnicity is not always asked for in such studies and, additionally, because ethnicity is a spectrum that can only be measured imperfectly, this is not always possible.

Reciprocally, research on the genetic structure of the human population has shown that when performing principal components analysis on an SNP genotyping dataset, a two-dimensional visual representation of the two first principal components allowed to clearly distinguish between groups from different geographical locations or with different origins [Novembre et al., 2008]. In other words, the first few principal components of the design matrix composed of the genetic covariates could well be good predictors of the individuals' origins or ethnicities. Hence, when no information is present in the data about the individuals' geographical origins or ethnicities, it may

be that this information could be imperfectly extracted from the genetic data using the first few principal components of the design matrix. And in fact, adjusting for the first few principal components is a widely used method to minimise spurious associations due to population stratification in such studies [Price et al., 2006].

A natural question that stems from this observation is to ask when and how one can hope to infer information about the potential confounders from the observed data and whether it is possible to use that information to remove some of the confounding-induced bias in the estimates. Following Ćevid et al. (2018), we show in the following sections that under some assumptions about the structure of the data and the confounders, the coefficient of interest β – which measures the effect of the observed covariates on the outcome – can be estimated with the same l_1-error rate as the Lasso in the case where there are no confounders.

3.1.1 Modelling assumptions

Linear confounding model Consider the linear confounding model M1. We assume in the following that the confounding variables are correlated with the predictors, where $X = Z\Gamma + E$ and Γ is such that the correlation between the rows of Z and E is null, i.e. that $\mathrm{Cov}(Z, E) = 0$. We write the model again below for clarity:

$$\begin{aligned} X &= Z\Gamma + E, \\ Y &= X\beta + Z\delta + \nu. \end{aligned} \tag{M1}$$

In this model, β describes the linear effect of the observed covariates X on the outcome Y and is generally assumed to be sparse. δ describes the linear effect of the confounding variables on the outcome Y, and Γ describes the linear effect of the confounding variables on the observed covariates X. Additionally, since we do not impose any restrictions on δ, the model does not change under the transformations $Z \leftarrow Z\mathrm{Cov}(Z)^{-1/2}$ and $\delta \leftarrow \mathrm{Cov}(Z)^{1/2}\delta$, and hence we can conveniently assume without loss of generality that $\mathrm{Cov}(Z) = I_r$.

Finally, we assume in the following that the rows of X and Z are independent, identically distributed, and jointly Gaussian, and that ν is a vector of sub-Gaussian errors with mean zero and standard deviation σ_ν that is independent of X and Z.

Linear perturbation model Both for interpretation and for deriving the following results, it will be helpful in the following to see that we can rewrite the model M1 as a linear model with a perturbed coefficient. Let Xb be the L_2 projection of $Z\delta$ onto X, so that we can write $Z\delta = Xb + (Z\delta - Xb)$ where $Z\delta - Xb$ is orthogonal to X (so $X^T(Z\delta - Xb) = 0$). Because Xb is the L_2 projection of $Z\delta$ onto X, we can write it as a linear predictor, and we have that:

$$\begin{aligned} Xb &= X\mathbb{E}[XX^T]^{-1}\mathbb{E}[X^T Z\delta] \\ &= X\mathrm{Cov}(X)^{-1}\mathbb{E}[(ZZ^T\Gamma)^T]\delta \\ &= X\mathrm{Cov}(X)^{-1}\Gamma^T\delta. \end{aligned}$$

Hence, by denoting $\Sigma = \mathrm{Cov}(X)$, we have $b = \Sigma^{-1}\Gamma^T\delta$. By inserting for Xb, we can then rewrite the linear confounding model M1 as:

$$Y = X(\beta + b) + (Z\delta - Xb) + \nu.$$

But since we assumed that the rows of X and Z were jointly gaussian, as we have that $Z\delta - Xb$ is uncorrelated with X by construction, we also know that $Z\delta - Xb$ and X are independent. Thus we can think of $Z\delta - Xb$ as an additional, uninformative error term in this linear model, and by writing $\epsilon = Z\delta - Xb + \nu$ for some new error term in this model, we can rewrite the model as:

$$Y = X(\beta + b) + \epsilon. \tag{M3}$$

Relationship between the two models By writing M1 in the form of M3, it is important to notice that we operate a decomposition of $Z\delta$ into the two terms Xb and $Z\delta - Xb$. We can think of the first of these terms as the part of the confounding effect that can be explained by the observed data X, and of the other as an additional effect on the outcome Y on which we don't have any information, and which we can thus think of as an additional error term in this linear model.

Furthermore, this rewriting gives some mathematical illustration of what we hypothesised with example E1 on genetic determinants of cardiovascular diseases: the key idea is to leverage the fact that the confounding effect can be partially explained with the observed covariates X.

Identification of β in the perturbed linear model Finally, writing M1 as M3 also illustrates that, in general, without additional hypotheses, β is not identifiable as it can't be separated from b, and we can only hope to estimate $\beta + b$ (which might as well be infeasible in the high-dimensional setting if b is dense).

However, we are primarily interested in β, which models the unconfounded effect of the observed covariates X on the outcome Y. Hence, we focus in the next sections on the conditions which, in this high-dimensional linear model, allow us to retrieve the coefficient β from the composite term $\beta + b$.

3.1.2 Elements of intuition on the use of spectral transformations

In example E1, we argued that, in some cases, one of the potential confounders in genetic studies could be imperfectly estimated from the first few right singular vectors of the observed data. This observation raises two questions:

(i) first, in what circumstances can some information about the confounders be retrieved from the first few right singular vectors of the design matrix (or equivalently from the top principal components of the sample covariance matrix)?

(ii) then, when this is the case, how can one attempt to adjust for confounding using only the observed data?

In this section, we present some illustrative intuitions on these questions using our modelling notations and present the class of spectral transformations introduced by Ćevid et al. (2018), which attempts to leverage these observations.

Alignment of the confounding bias b In model M3, recall that b models the bias due to the confounding effect, and Xb is the part of the confounding that can be explained from the observed covariates X. Using these notations, we can reformulate the first question into asking why b would approximately lie in the span of the first few right singular vectors of the design matrix X.

With a single confounder ($r = 1$) To illustrate this, we first place ourselves in the case of a linear confounding model with a single confounder (so $r = 1$). We show that in this setting, under mild assumptions, Γ^T will be well aligned with the top right singular vector of the design matrix, and then that b is aligned with Γ^T.

Recall that in the linear confounding model M1, we had $X = Z\Gamma + E$ where we assumed $\mathrm{Cov}(Z, E) = 0$ and where we also assumed without loss of generality that $\mathrm{Cov}(Z) = I_r$. Hence, by writing $\Sigma_E = \mathrm{Cov}(E)$ we have:

$$\Sigma = \mathrm{Cov}(X) = \mathrm{Cov}(Z\Gamma + E, Z\Gamma + E) = \Gamma^T\Gamma + \Sigma_E.$$

Let $\Sigma_E = I_p$ (so the unconfounded design matrix has independent covariates). Notice then that $\Gamma^T \in \mathbb{R}^p$ is a column vector, where we can write:

$$\Gamma^T\Gamma = \|\Gamma^T\|_2^2 \left(\frac{1}{\|\Gamma^T\|_2}\Gamma\right)^T \left(\frac{1}{\|\Gamma^T\|_2}\Gamma\right) := \theta\gamma\gamma^T \quad \text{where} \quad \begin{array}{l} \theta = \|\Gamma^T\|_2^2 \\ \gamma = \Gamma^T \end{array}$$

By construction, $\gamma \in \mathbb{R}^p$ is a unit column vector, and we can recognise that the covariance matrix $\Sigma = I_p + \theta\gamma\gamma^T$ is the covariance matrix of a spiked covariance model. Thus, the eigenvalues of Σ are known and given by $1 + \theta, 1, ..., 1$, where the top eigenvector is γ, with eigenvalue $1 + \theta$.

By assumption, X is sampled i.i.d. from a distribution with covariance matrix Σ. Thus, if the columns of X are centred, the sample covariance matrix of X is given by:

$$\hat{\Sigma} = \frac{1}{n}X^TX.$$

By writing the singular value decomposition of X as $X = UDV^T$, we have that:

$$\hat{\Sigma} = \frac{1}{n}VDU^TUDV = \frac{1}{n}VD^2V^T.$$

Hence, when X is centred, the first right-singular vector of X is equal to the top eigenvector of the sample covariance matrix $\hat{\Sigma}$. Because we know that the first eigenvector of the true covariance matrix Σ is γ, we may therefore investigate when the first right-singular vector of X is likely to be well aligned with γ.

Let $\hat{\gamma}$ denote the first right singular vector of X which, as we have seen, is also the first eigenvector of the sample covariance matrix $\hat{\Sigma}$.

When $p < n$, and when the rows of X are sub-Gaussian, we already know that $\hat{\gamma}$ and γ will be close, in the sense that, with high probability, $\min_{\alpha \in \{-1,1\}} \|\gamma - \alpha\hat{\gamma}\|$ will be upper-bounded by a term which converges to zero as n increases (see Chapter 3 of the lecture notes in Modern Statistical Methods).

With high-dimensionality ($p > n$), however, this result does not hold without additional assumptions. Nevertheless, with the same setting, the literature on principal components analysis has shown that, in a spiked covariance model like this one, under certain conditions, γ and $\hat{\gamma}$ are well aligned asymptotically. As illustrated in Johnstone and Paul (2018), when $p/n \to \eta$ with $\eta > 1$:

(i) asymptotically, if $\theta > \sqrt{\eta}$ (value known as a phase transition), the top singular value of X will with high probability be larger than the rest of the singular values

(ii) asymptotically, if $\theta > \sqrt{\eta}$, the inner product of the first right singular vector of X with the first eigenvector of Σ will be upper-bounded, with:

$$|\gamma^T \hat{\gamma}| \to \frac{1 - \eta/\theta^2}{1 + \eta/\theta}.$$

We do not prove these statements in this essay, but a rigorous proof that extends to the setting with multiple spikes can be found in Theorem 4 from Paul (2007). In our case, to summarise, for a given asymptotic regime η and with θ large enough, one may hope that the first right singular vector of X and the first eigenvector of Σ are in the same direction, in other words, are well aligned.

We now show that b and γ are aligned. Using our derivation of b from the last section, noticing that $\delta \in \mathbb{R}$ in our setting, and using the Sherman-Morrison formula, we can write b as follows:

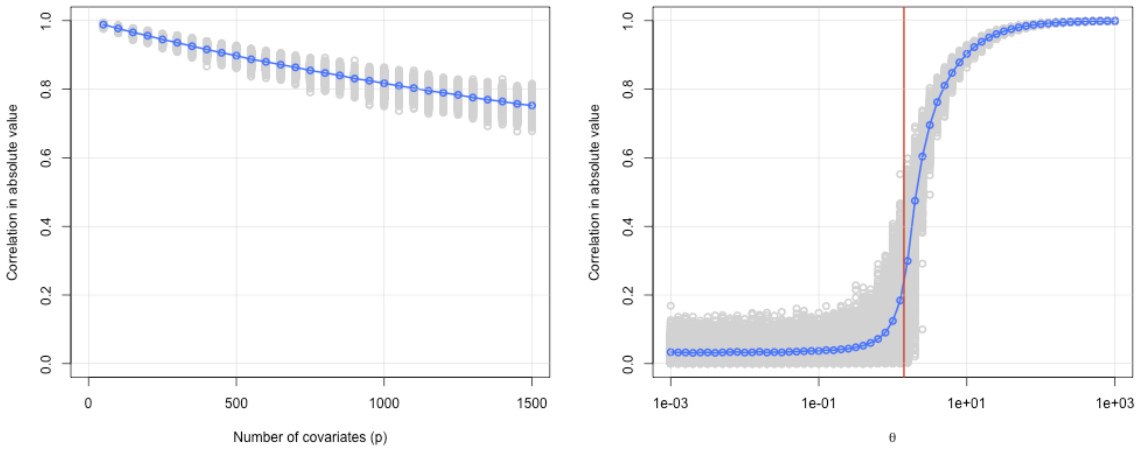
$$\begin{aligned}
b &= \Sigma^{-1}\Gamma^T\delta \\
&= \frac{\delta}{\sqrt{\theta}}(I_p + \theta\gamma\gamma^T)^{-1}\gamma \\
&= \frac{\delta}{\sqrt{\theta}}(I_p - \theta(1 + \theta\gamma^T\gamma)^{-1}\gamma\gamma^T)\gamma \\
&= \frac{\delta}{\sqrt{\theta}}(I_p - \frac{\theta}{1+\theta}\gamma\gamma^T)\gamma \\
&= \frac{\delta}{\sqrt{\theta}(1+\theta)}\gamma.
\end{aligned}$$

Putting the pieces together, in this setting, we see that b the bias due to the confounding is exactly proportional (and hence aligned) to γ, which, itself, is likely to be well aligned with the top right singular vector of X under the right assumptions.

Thus, in this illustrative example, asymptotically, and when $\theta = \|\Gamma^T\|_2^2$ is large enough, b will be well aligned with γ, and hence will also be well aligned with the first right singular vector of the design matrix of the observed data X.

To quickly illustrate this in practice, we present here some empirical results from simulated data. We generate the data from the linear confounding model M1 with a single confounder (so $r = 1$), where $\Sigma_E = I_p$ and $\gamma \in \mathbb{R}^p$ is such that $\gamma = \Gamma^T/\|\Gamma^T\|_2$ where the entries of Γ^T are randomly sampled from a standard normal distribution. Then, we take $\Sigma = I_p + \theta\gamma\gamma^T$ for $\theta \in \mathbb{R}$ and randomly sample the rows of X from a multivariate normal distribution with mean zero and covariance matrix Σ. All results are based on $N = 1000$ independent simulations.

In Figure 4a, we generate the data with $n = 300$ and with p varying from 50 to 1500, where $\theta = 8$. In this setting, we see that the correlation between b and the first right singular vector of the design matrix X remains high (more than 75%). Notice that in this case, while θ is much larger than $\sqrt{p/n}$, it is relatively low compared to $\|\Gamma^T\|_2^2 \approx p\text{Var}(\Gamma^T)$. If we used $\Sigma = I_p + \Gamma^T\Gamma$ for the covariance matrix instead, θ would be equal to $\|\Gamma^T\|_2^2$ which would be close to p with high probability, and the correlations of Figure 4a would be much higher. In Figure 4b, we generate data with $n = 300$, $p = 600$ and with θ varying from 10^{-3} to 10^3. We observe that the correlation between b and the first right singular vector of the design matrix starts to become high around the asymptotic limit value $\theta = \sqrt{p/n} = \sqrt{2}$ (red line in the graph).



(a) (b)

Figure 4: Dependence of the correlation between b and the first right singular vector of the design matrix X depending (a) on the number of covariates, and (b) on θ

In this example, we see again that given the asymptotic regime η, if θ is large enough, b will be well aligned with the first right singular vector of X. Now, if Γ^T is distributed from a normal distribution with mean zero and variance ϑ^2, we know that $\|\Gamma^T\|_2^2$ will be close to $p\vartheta^2$. But since we can write $\Gamma^T\Gamma = \|\Gamma^T\|_2^2\gamma\gamma^T$ where $\gamma = (1/\|\Gamma^T\|_2)\Gamma^T$ and $\|\gamma\|_2 = 1$, assuming that we are close to the asymptotic setting,

it follows that if ϑ^2 is much larger than $\frac{1}{p}\sqrt{\frac{p}{n}}$ (or equivalently if p if larger than $\frac{1}{n\vartheta^4}$), then b will likely be well aligned with the first right singular vector of X, and it makes sense to use a spectral transformation that shrinks in the direction of γ.

With multiple confounders ($r > 1$) We do not attempt to illustrate how this may hold in a more general setting with more than one confounder. However, it makes intuitive sense to hope that, under some distributional assumptions, the confounders will also be responsible for the largest singular values of Σ, whose singular vectors we expect to be well aligned with the top few right-singular values of X. Thus the confounding bias b will lie approximately in the span of the first few right singular vectors corresponding to the largest singular values of the observed design matrix X.

Use of spectral transformations Building on this, we may now want to exploit the fact that b will approximately lie in the span of the first few right singular vectors of X to adjust for some of the confounding bias. Notice first that the perturbed linear model M3 can be rewritten as:

$$Y = X\beta + Xb + \epsilon,$$

and so for any transformation matrix $F \in \mathbb{R}^{n\times n}$, we can write:

$$FY = FX\beta + FXb + F\epsilon. \tag{M4}$$

Ideally, we would like to find a transformation matrix F such that FXb will be much smaller than Xb in some norm, but where $FX\beta$ remains approximately equal to $X\beta$, or in other words, we wish to shrink the signal along the direction of b, but not β. But as we argued that b lied approximately in the span of the first few right singular vectors of X, it makes sense to try to find a transformation F that will shrink the signal in the direction of those first few singular vectors, in the hope that β is not too well aligned with these singular vectors. These linear transformations are called spectral transformations. Putting the pieces together, under the additional assumption that the error term $F\epsilon$ is well behaved, it then makes sense to use a usual high-dimensional regression like Lasso to regress FX on FY to estimate β.

Class of spectral transformations Following Ćevid et al. (2018), we introduce a class of multiple spectral transformations that lower the contribution of the first few singular vectors to the design matrix, and present the general idea of the method for estimating β using these spectral transformations.

In the setting of a high-dimensional linear regression with $n << p$, let the singular value decomposition of the design matrix X be $X = UDV^T$, with $U \in \mathbb{R}^{n\times n}$ a matrix whose columns are the left singular vectors of X, $D \in \mathbb{R}^{n\times n}$ a diagonal matrix whose diagonal entries are the singular values of X, and $V \in \mathbb{R}^{p\times n}$. Then, the class of spectral transformations considered by Ćevid et al. (2018) is that of the ones that shrink the singular values of X while keeping the same matrices U and V. Since U is semi-unitary, this can be done by taking a transformation F with:

$$F = UQU^T \quad \text{where} \quad Q = \begin{pmatrix} \tilde{d}_1/d_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2/d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_n/d_n \end{pmatrix}.$$

Here, $d_1, d_2, ..., d_n$ are the singular values of the design matrix X, i.e. the diagonal entries of D, and $\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_n$ are new singular values that are chosen such that $\forall i \in \{1, ..., n\} : \tilde{d}_i \leq d_i$ so that the spectral transformation is shrinking the singular values of the design matrix.

Then, we can write the transformed design matrix $\tilde{X}$ as:

$$\tilde{X} = FX = UQU^T UDV^T = U \begin{pmatrix} \tilde{d}_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_n \end{pmatrix} V^T.$$

In this rather general setting, multiple particular spectral transformations can be defined depending on the values taken by $\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_n$. In this essay, we focus solely on PCA adjustment and Trim transform, as introduced by Ćevid et al. (2018).

PCA adjustment Following from the intuition given by example E1 on the genetic determinants of cardiovascular diseases, a seemingly natural procedure would be to regress out the first few principal components from the design matrix before applying a usual high-dimensional regression model like Lasso. Notice that this bears some similarity with the idea of using factor models to adjust for confounding since we can think of the first few right singular vectors as surrogate covariates for the confounders. Furthermore, adjusting for the first few principal components is equivalent to directly removing the contributions of the first few singular vectors to the design matrix, and hence this is equivalent to defining:

$$\forall i \in \{1, ..., n\} : \tilde{d}_i = d_i \mathbb{1}_{i>k} \quad \text{for} \quad k \in \{1, ..., n\}.$$

Despite it being seemingly more natural, PCA adjustment has two major drawbacks in that one needs to know in advance the number of principal components r to adjust for (which in practice may be difficult to determine) and that, by removing all of the contributions of the first r principal components of the design matrix, one also risks to remove part of signal $X\beta$ which we want to estimate.

Trim transform Building in part on these two limitations, Ćevid et al. (2018) propose to use another spectral transformation known as Trim transform, which limits all singular values to a constant τ which can be chosen, or inferred from the data. This amounts to defining:

$$\forall i \in \{1, ..., n\} : \tilde{d}_i = \max(\tau, d_i).$$

Here τ can, for instance, be taken to be the median of all singular values: $\tau = d_{\lfloor n/2 \rfloor}$.

Estimating procedure Finally, for a given spectral transformation F defined as above, the resulting estimating procedure suggested by Ćevid et al. (2018) estimates β by using a Lasso regression of the transformed design matrix $\tilde{X} = FX$ on the outcome $\tilde{Y} = FY$, so that:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda \|\beta\|_1. \tag{P1}$$

3.1.3 Point parameter estimation of β

In this section, we present the main result derived in Ćevid et al. (2018) regarding point parameter estimation of β using the estimating procedure P1, and in particular, the derivation of an upper bound on the l_1-estimation error rate in the linear confounding model. To do so, we start by defining some additional notations.

Additional notations Let S be the support set of β and s be the size of S (so that only s coordinates of β are not null). Then, for any rectangular matrix A, let $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ $\lambda_i(A)$ be the maximum, minimum and i-th singular values respectively. Let also $\tilde{\Sigma}$ be the empirical covariance matrix of the transformed design matrix $\tilde{X}$ and $\hat{\Sigma}$ be the empirical covariance matrix of the original design matrix X, such that:

$$\hat{\Sigma} = \frac{1}{n} X X^T \quad \text{and} \quad \tilde{\Sigma} = \frac{1}{n} \tilde{X} \tilde{X}^T.$$

Finally, recall that in the derivation of the improved upper bound of the l_1-estimation error for the Lasso with no confounding, one key step and assumption is the validity of the compatibility condition, which states that the compatibility constant is not null [Van de Geer, 2007] (or see lecture notes in Modern Statistical Methods). We define analogously some compatibility constant in this setting. Let ϕ_M denote this quantity for any matrix $M \in \mathbb{R}^{n \times n}$, such that:

$$\phi_M = \inf_{\|\alpha\|_1 \leq 5\|\alpha_S\|_1} \frac{\sqrt{\alpha^T M \alpha}}{\frac{1}{\sqrt{s}} \|\alpha_S\|_1}.$$

l_1-error estimation rate in the perturbed linear model To derive our final result, we first obtain an upper bound for the l_1-error estimation rate for the perturbed linear model M3 estimated by the procedure P1, and then make the link with the linear confounding model we are most interested in. We see that this upper bound depends on the l_2-norm of $\tilde{X}b$, on the compatibility constant for the transformed matrix $\tilde{\Sigma}$, as well as on constants in the following theorem.

Theorem 3.1. *Consider the model M3 with* $\max_i \Sigma_{i,i} = \mathcal{O}(1)$ *where* $\Sigma = \mathrm{Cov}(X)$, *and where the error term* $\epsilon \in \mathbb{R}^n$ *is a vector of sub-Gaussian errors independent of* X *with standard deviation* σ. *Let* $F \in \mathbb{R}^{n \times n}$ *be an arbitrary linear transformation. Then, for the procedure P1 with transformation matrix* F, *let the penalty level* λ *be:*

$$\lambda = A\sigma\sqrt{\frac{\log p}{n}}\lambda_{\max}(F)^2.$$

Then, with probability at least $1 - 2p^{1-A^2/(32\max_i \Sigma_{i,i})} - pe^{-n/136}$, *we have:*

$$\|\hat{\beta} - \beta\|_1 \leq C_1 \frac{s\lambda}{\phi_{\tilde{\Sigma}}} + C_2 \frac{\|\tilde{X}b\|_2^2}{n\lambda}$$

where C_1 *and* C_2 *are constants depending only on* A.

The proof of this theorem uses roughly the same set of main arguments as the proof in the setting with no confounding (see [Van de Geer, 2007] or lecture notes in Modern Statistical Methods), and can be found in Ćevid et al. (2018). It is not presented here.

It is worth noting that Theorem 3.1 highlights two important dependencies of the upper bound on the l_1-estimation error in this setting:

(i) first, as we had already intuited, one key component in getting accurate estimates of β is to be able to make $\|\tilde{X}b\|_2^2$ small, that is, to shrink the signal Xb which is the part of the confounding effect that is correlated with X

(ii) secondly, we can see that the upper bound is inversely related to $\phi_{\tilde{\Sigma}}$, where, as $\tilde{\Sigma}$ is positive semi-definite, one can think that a possible setting that would make $\phi_{\tilde{\Sigma}}$ close to zero (and hence increase the upper bound) would be when the singular values of $\tilde{\Sigma}$ are too small, thus when the transformation F has shrunk the singular values of X too much

Building on this, and following Ćevid et al. (2018), we derive in Corollary 3.1.1 an upper bound for the l_1-estimation error rate of β under some assumptions in close relationship with the two dependencies highlighted above.

Corollary 3.1.1. *Consider the model M3 with* $\max_i \Sigma_{i,i} = \mathcal{O}(1)$ *where* $\Sigma = \mathrm{Cov}(X)$, *and where the error term* $\epsilon \in \mathbb{R}^n$ *is a vector of sub-Gaussian errors independent of* X *with standard deviation* σ. *Suppose that* $\lambda_{\min}(\Sigma)$ *is bounded away from zero. For the perturbation coefficient* b, *assume that:*

(i) $\|b\|_2^2 = \mathcal{O}\left(\frac{s\sigma^2 \log p}{p}\right)$.

Assume additionally that the linear transformation matrix F *in the procedure P1 verifies* $\lambda_{\max}(F) = 1$, *and that*

(ii) $\lambda_{\max}(\tilde{X}) = \mathcal{O}_p(\sqrt{p})$,

(iii) $\phi_{\tilde{\Sigma}} = \Omega_p(\lambda_{\min}(\Sigma))$.

Then, for a penalty level λ with asymptotic rate $\sigma\sqrt{\frac{\log p}{n}}$, the l_1-estimation error has the following rate:

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

Proof. Recall that, from Theorem 3.1, we have:

$$\|\hat{\beta} - \beta\|_1 \leq C_1\frac{s\lambda}{\phi_{\tilde{\Sigma}}} + C_2\frac{\|\tilde{X}b\|_2^2}{n\lambda}$$

Then, using (i) and (ii), and letting the singular value decomposition of $\tilde{X}$ be such that $\tilde{X} = \tilde{U}\tilde{D}\tilde{V}^T$ where we have $\tilde{U} \in \mathbb{R}^{n\times n}$, $\tilde{D} \in \mathbb{R}^{n\times n}$ and $\tilde{V} \in \mathbb{R}^{p\times n}$, we have:

$$\|\tilde{X}b\|_2^2 = b^T\tilde{X}^T\tilde{X}b = b^T\tilde{V}\tilde{D}^2\tilde{V}^Tb \leq \lambda_{\max}(\tilde{X})^2\|b\|_2^2 = \mathcal{O}\left(s\sigma^2\log p\right).$$

Thus, since λ has asymptotic rate $\sigma\sqrt{\frac{\log p}{n}}$, it follows that:

$$\frac{\|\tilde{X}b\|_2^2}{n\lambda} = \mathcal{O}_p\left(\frac{s\sigma^2\log p\sqrt{n}}{n\sigma\sqrt{\log p}}\right) = \mathcal{O}_p\left(s\sigma\sqrt{\frac{\log p}{n}}\right).$$

Furthermore, using (iii), we have that:

$$\frac{s\lambda}{\phi_{\tilde{\Sigma}}} = \mathcal{O}_p\left(\frac{s\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

Hence, using the upper bound from in Theorem 3.1, we obtain:

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

□

l_1-estimation error rate for the linear confounding model Following this derivation, by relating the assumptions of the two models (whose relationships we do not detail here), the authors derive the following upper bound on the l_1-estimation error rate. The full derivation is done by Ćevid et al. (2018) (Lemma 1). Notice that the assumptions of Theorem 3.2 do not depend on b, whose exact behaviour is difficult to characterise and which may be hard to interpret in practice, but on the parameters of the original linear confounding model M1.

Theorem 3.2. *Consider the linear confounding model M1 with $\max_i \Sigma_{i,i} = \mathcal{O}(1)$ where $\Sigma = \mathrm{Cov}(X)$, and where $\frac{\lambda_{\max}(\Sigma_E)}{\lambda_{\min}(\Sigma_E)} = \mathcal{O}(1)$. Suppose that $\lambda_{\min}(\Sigma)$ is bounded away from zero. Assume additionally that:*

(i) $\lambda_{\min}(\Gamma) = \Omega(\sqrt{p})$.

Assume additionally that the linear transformation matrix F in the procedure P1 verifies $\lambda_{\max}(F) = 1$, and that

(ii) $\lambda_{\max}(\tilde{X}) = \mathcal{O}_p(\sqrt{p})$,

(iii) $\phi_{\tilde{\Sigma}} = \Omega_p(\lambda_{\min}(\Sigma))$.

Then, for a penalty level λ with asymptotic rate $\sigma\sqrt{\frac{\log p}{n}}$, the l_1-estimation error has the following rate:

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

From this theorem, we can see that under the modelling assumptions of model M1 and some additional assumptions, the estimating procedure P1 provides an asymptotic upper bound on the l_1-estimation error rate of the coefficient of interest β which achieves the same l_1-error rate as the usual Lasso with no confounding.

A natural question that stems from this theorem is then to ask when the assumptions of Theorem 3.2 are verified, and in what cases they may fail to hold.

The first assumption (assumption (i)) in Theorem 3.2 is related to the denseness of the confounding in the sense that each confounding variable is correlated with many predictors, thus that the rows of Γ are dense in a certain sense. This is the case, for instance, if Γ is drawn at random with rows distributed i.i.d. and sub-Gaussian. This is shown in detail by Ćevid et al. (2018) (Lemma 1).

More intuitively perhaps, in the special case where the linear confounding model is such that $r = 1$ (thus when there is a single confounder), we can link this assumption to our discussion from Section 3.1.2. Notice that when $\Gamma \in \mathbb{R}^{1\times p}$, the only singular value of Γ is equal to $\|\Gamma^T\|_2$ because we only require the singular vector to have unit norm. Thus, in this case, assumption (i) reduces to $\|\Gamma^T\|_2 = \Omega(\sqrt{p})$. This directly echoes what we argued in Section 3.1.2, as we stated that we needed $\|\Gamma^T\|_2$ to be much larger than $\sqrt{p/n}$ for the correlation between b and the first right singular vector of the design matrix X to be high (which we need for the spectral transformation to adjust for the confounding bias correctly).

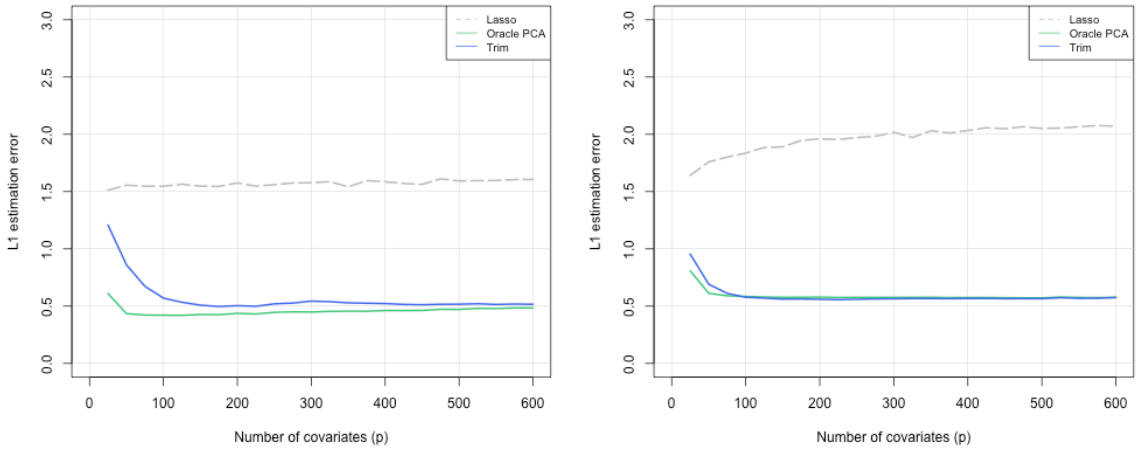
The second and third assumptions ((ii) and (iii) in the statement of the theorem) are more closely related to the choice of spectral transformation F. First, the second assumption states that the largest singular value of the transformed design matrix $\tilde{X}$ is at most of order $\sqrt{p}$. Since F shrinks the singular values of the design matrix X, this assumption is directly verified if the largest singular value of X is of order $\sqrt{p}$. When this is not the case, the assumption requires that the first few singular values of X are shrunk enough by F so that the new transformed singular values $\tilde{d}_1, ..., \tilde{d}_n$ are at most of order $\sqrt{p}$. The following lemma, whose proof is derived in full by Ćevid et al. (2018) (Lemma 2), but which we do not prove here, shows that this is true under some distributional assumptions for both PCA adjustment with r adjusted principal values, and for the Trim transform if $\lfloor \frac{n}{2} \rfloor > r$.

Lemma 3.3. *Assume that $p > n$ and that X has i.i.d. sub-Gaussian rows with covariance matrix $\Sigma = \Gamma^T\Gamma + \Sigma_E$, where $\Gamma \in \mathbb{R}^{r\times p}$ and $\lambda_{\max}(\Sigma_E) = \mathcal{O}(1)$. Then, we have $d_{r+1} = \mathcal{O}(\sqrt{p})$.*

Finally, the third assumption states that the compatibility constant $\phi_{\tilde{\Sigma}}$ is at least of the same order as the smallest singular value of Σ. This is verified if the spectral transformation F does not shrink the signal $X\beta$ too much (i.e. by shrinking in the direction of the confounding bias, we do not want to remove $X\beta$ the signal of interest as well). We do not state the result in full here, but Ćevid et al. (2018) show that the assumption is verified for PCA adjustment with r adjusted principal values when r is fixed, $\frac{1}{p}\sum_{i,j=1}^{p}|\Sigma_{E_{i,j}}|$ is upper-bounded, and $\frac{sp\log p}{n^2} \to 0$.

Simulations To illustrate these results, we briefly present some empirical results from simulated data. We generate the data from the linear confounding model M1, with $n = 300$, $r = 6$ and with β such that only the first 6 entries are non-zero. The entries in Γ, δ and ν are randomly sampled from the standard normal distribution. Finally, the rows of E are randomly sampled from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_E = \sigma_E^2 I_p$ where $\sigma_E = 2$, and the rows of Z are randomly sampled from a multivariate normal distribution with mean zero and covariance matrix I_r. All results are based on $N = 1000$ independent simulations.

We observe in Figure 5, as do Ćevid et al. (2018), that, no matter whether the penalty term in the Lasso regression of the transformed design matrix $\tilde{X}$ on the transformed outcome $\tilde{Y}$ is chosen equal to its theoretical rate $\sqrt{\log p/n}$ or via cross-validation, both the Trim transform and Oracle PCA perform better than using the Lasso without accounting for confounding.



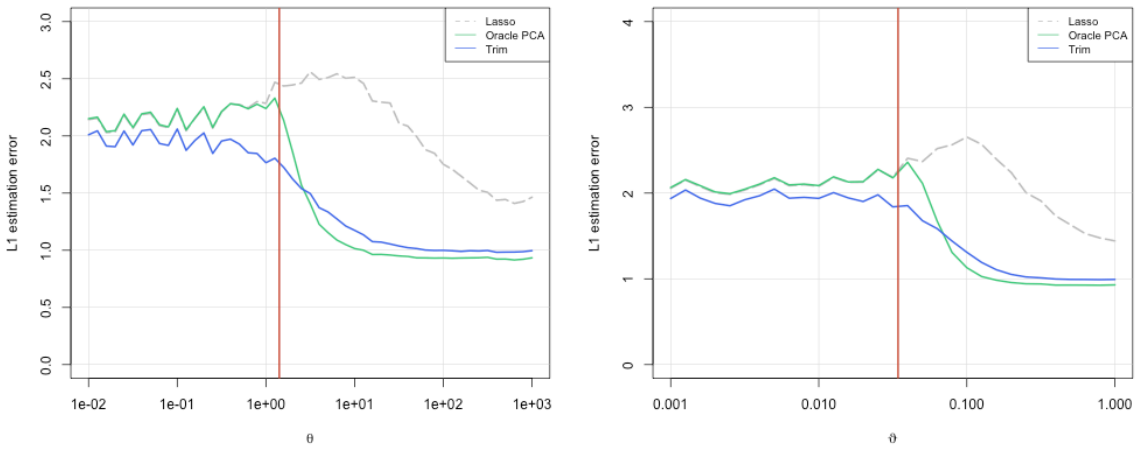
(a) (b)

Figure 5: Dependence of the l_1-estimation error on the number of predictors p depending on whether the penalty term is chosen (a) equal to its theoretical optimal value, or (b) using cross-validation

Finally, we push a little further our discussion in the case where the data is simulated with the linear confounding model M1 but with a single confounder, where $\Sigma_E = I_p$ (so $\sigma_E = 1$) and $\gamma \in \mathbb{R}^p$ is such that $\gamma = \Gamma^T / \|\Gamma^T\|_2$ where the entries of Γ^T are randomly sampled from a standard normal distribution. To have control over the spiked covariance model from which X is simulated, we then take $\Sigma = I_p + \theta\gamma\gamma^T$ for $\theta \in \mathbb{R}$ and randomly sample the rows of X from a multivariate normal distribution with mean zero and covariance matrix Σ.

Recall that we argued in Section 3.1.2 that the confounding bias b from the perturbed linear model M3 would only be well aligned with the first right singular vector of X if θ was large enough (asymptotically, when $\theta > \sqrt{\eta}$ where $p/n \to \eta$). Hence, we may expect the procedure studied here to only work for large enough values of θ. In fact, we observe in Figure 6a that when θ is small enough (and since $p = 600$ and $n = 300$ here, roughly when $\theta < \sqrt{600/300} = \sqrt{2}$, which is highlighted by the red line in the graph), the Lasso after applying a spectral transformation does not perform better than the Lasso without adjusting for confounding. This is analogous to the assumption (i) in Theorem 3.1 as a large enough θ will guarantee that the (only) singular value of Γ^T is large enough. Finally, for completeness, notice that one may argue as well that when θ is very small, the contribution of the confounders to the design matrix X becomes very small too, which implies that we are straying further from the setting of confounded data.

Using our notations from Section 3.1.2, we also investigate the dependence of the l_1-estimation error on ϑ when the entries of Γ^T are randomly sampled from a normal distribution with mean zero and covariance ϑ^2. We observe in Figure 6b that the PCA adjustment and Trim transform perform better only when ϑ is larger than the theoretical asymptotic rate $(pn)^{-1/4} \approx 0.034$ (red line in the graph), where b will start to be aligned with the first right singular vector of the design matrix X.



(a) (b)

Figure 6: Dependence of the l_1-estimation error (a) on θ and (b) on ϑ

3.1.4 Case study: the determinants of tourism in Germany

This section illustrates with a small case study how one may try to adjust for potential confounding bias in practice. We try to replicate our introductory example E2 on the determinants of tourism. We use data on German cities and urban areas from the European Data Portal (`https://ec.europa.eu/eurostat/`) that were collected in 2011 and which record many characteristics ranging over broad themes covering demographics, economic activity, environment, urban planning, etc.

We consider $n = 101$ cities, for which $p = 206$ continuous covariates have been measured. The full list of covariates and the dataset can be found in the GitHub repository. While Ćevid et al. (2018) do not mention any strict limit cases for using spectral transformations, we acknowledge that our sample size and number of covariates are at the lower limit case for applying such methods in practice.

The problem considered is that of determining covariates that are potentially associated with tourism in Germany. The outcome variable is the number of nights spent in tourist accommodation per 1000 inhabitants and year. Because the driving forces of tourism are complex, we do not remove covariates or impose any prior knowledge on the data. However, we allow ourselves to comment on the results obtained and acknowledge that doing so bears some inherent subjectivity.

Lasso regression without adjusting for confounding We start by using a Lasso regression with cross-validation on the observed (standardised) data without using any additional method to adjust for potential confounding. The resulting coefficient estimates $\hat{\beta}_{\texttt{CODE}}$ for the covariates with non-zero coefficient are presented in Table 1.

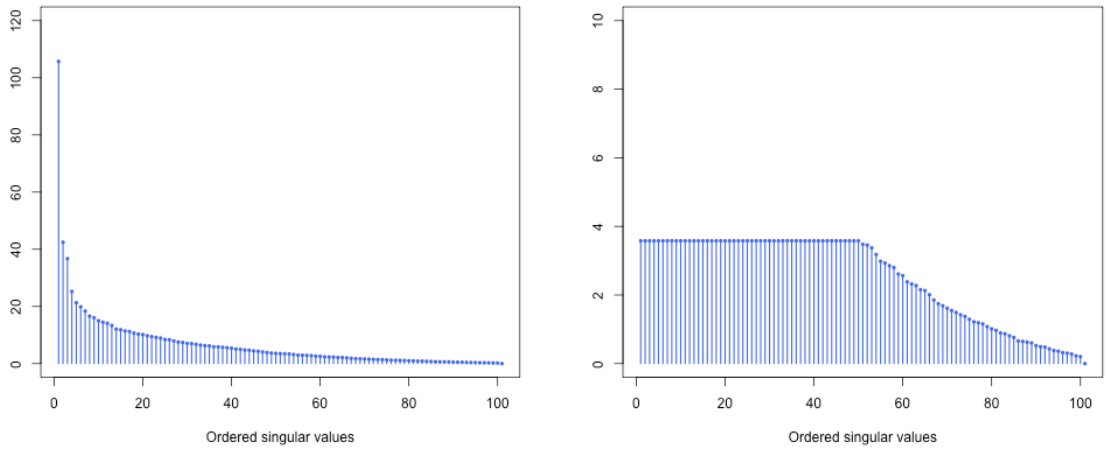
CODE	Covariate	$\hat{\beta}_{\texttt{CODE}}$
LAT	Latitude	0.0384
CR1003I	Number of cinema seats per 1000 residents	0.0306
CR1007V	Number of museum visitors (per year)	0.0044
DE1003I	Number of women per 100 men	0.0062
DE1061I	Population change over last year	0.0031
DE1073I	Proportion of population aged 25-34 years	0.1034
DE3002I	Proportion of households that are 1-person households	0.0815
EC2034V	Employment in financial and insurance activities	0.1085
SA1051V	Average price for buying an apartment	0.1145
TE2028I	Prop. of working age population qualified at level 3 or 4 ISCED	-0.0274
TE2031I	Prop. of working age population qualified at level 5 or 6 ISCED	0.2906

Table 1: Covariates with non-zero coefficients (and respective coefficients) in a Lasso regression without adjusting for confounding

For some of the remaining variables, the link between the covariate of interest and tourism is rather evident (for instance, the number of museum visitors per year). However, for others, the association appears as much less justifiable, and one may argue that they hide a confounding factor. For instance, it is rather unlikely that tourists are particularly keen on visiting cities with better-educated inhabitants. On the other hand, the fact that the population is relatively young and well educated look like good proxies or predictors of the fact that a city is wealthy and attractive, which may as well be one of the main reasons of its touristic appeal. In this setting, the unmeasured and ill-defined attractiveness of a city could be seen as a potential confounder, as it is likely to be associated with both younger, better-educated and wealthier populations, and with the number of tourists visiting every year. As a result, we may think that some of these associations would disappear had additional variables been measured, and adjusted for.

Notice finally that, for some remaining covariates in Table 1, like the latitude or the average price for buying an apartment, thinking of intuitive potential confounders is more difficult, although a direct link between tourism and these covariates is not clear either. Elucidating these associations would require further investigations, but one may argue that they are due to either the particularities of Germany (as many touristic German cities used to be in the Hanseatic League and are close to the North Sea), or reflect mediated associations (for instance as tourists' visits increase the demand for housing, which in turn is likely to increase the price of houses and apartments).

Lasso regression after spectral transformation Building on this, we applied the method described by Ćevid et al. (2018), and then used a Lasso regression on the transformed design matrix where the spectral transformation was set to be the Trim transform. We plot the singular values of the original design matrix in Figure 7a and of the transformed one in Figure 7b for illustration (notice the change of scale).



(a) (b)

Figure 7: Singular values (a) of the original (standardised) design matrix and (b) of the (standardised) design matrix after applying the Trim transform

The new coefficient estimates $\hat{\beta}_{\texttt{CODE}}$ for the covariates with non-zero coefficient are presented in Table 2. We observe that more parameters are shrunk to zero than in the first model. We argue that the associations of the outcome with the remaining covariates are less likely to be due to confounding. For instance, in the first model, keeping all other covariates equal, it would be difficult to argue that an increase in tourists' visits leads to an increase in employment in the financial and insurance sectors. However, when looking at the new model coefficients, it is easier to believe that an increase in tourists visits will lead to an increase in employment in trade, transport, hotels and restaurants, as tourists' visits directly affect these sectors.

CODE	Covariate	$\hat{\beta}_{\texttt{CODE}}$
CR1007V	Number of museum visitors (per year)	0.3280
DE2007I	Foreigners as a proportion of population	-0.0171
EC2032V	Employment in trade, transport, hotels, restaurants	0.1032
SA1050V	Average price for buying a house	0.0116

Table 2: Covariates with non-zero coefficients (and respective coefficients) in a Lasso regression after adjusting for confounding using Trim transform

In short, we can observe a positive association between tourism and the number of museum visitors per year; employment in trade, transport, hotels and restaurants; and the average price for buying a house, all of which appear as justifiable. We also observe a negative association between tourism and the proportion of foreigners, which is less clear but whose coefficient is also relatively small. Here, these coefficients are biased because of both high-dimensionality and the adjustment for confounding, and one would need to correct for that bias to interpret their values numerically. Although we do not apply it here, this could be done following the approach of Guo et al. (2020).

As a final and more general note, one may also argue, seeing these results, that the determinants of tourism are hard to grasp with such a dataset. Indeed, a city may drive many tourists because of its unique history, its beautiful architecture, or its annual festivals or celebrations, all of which can difficulty be measured.

3.2 Non-linear approaches

So far, we focused on a linear confounding model and did not assume any structural relationships between the covariates and the outcome. This has two significant drawbacks: first, it implies that the estimated coefficients β cannot be interpreted as causal effects, and then, a linear model may be an incorrect specification of the relationships.

Recently, Wang and Blei (2019) presented a novel approach that attempts to adjust for confounding in a non-linear and causal framework using unsupervised machine learning and predictive model checking. Recall from Section 2.2.1 that, using counterfactual notations, we wish to determine $\mathbb{E}[Y(x)]$ for $x \in \mathbb{R}^p$ where $Y(x)$ is the counterfactual outcome had the covariates X been set artificially to x.

In this setting, we still think of the confounders Z as unobserved variables, and we would like to identify the distribution $f(Y(x))$, which depends on the unobservable distributions $f(Z = z)$ and $f(Z = z, X = x)$ in the following way:

$$f(Y(x) = y) = \int_{\mathcal{Z}} f(Y = y \mid X = x, Z = z) f(Z = z) \, \mathrm{d}z,$$

$$\text{where} \quad f(Z = z) = \int_{\mathcal{X}} f(Z = z \mid X = x) f(X = x) \, \mathrm{d}x = \int_{\mathcal{X}} f(Z = z, X = x) \, \mathrm{d}x.$$

We do not present the approach of Wang and Blei (2019) in detail in this essay but only state the main idea. In short, the method uses a probabilistic factor model for $f(X = x, Z = z)$ to estimate surrogate variables for the confounders. Then, if the distribution $f(X = x, Z = z)$ is known, the marginal distribution $f(Z = z)$ and the conditional one $f(Z = z \mid X = x)$ can be computed, and $f(Y(x))$ retrieved.

In contrast with the regular factor models (as introduced in Section 2.2.3), a probabilistic factor model estimates the latent variable using random variables rather than fixed quantities. One such model is probabilistic principal components analysis, which was introduced by Tipping and Bishop (1999), who give an overview of how the model is fitted and the practical cases where it can be applied.

Probabilistic PCA is just one example of probabilistic factor models, and the approach of Wang and Blei (2019) builds on fitting multiple models and choosing the one with the best results. To do so, the method checks for the predictive power of each probabilistic factor model, selects the best depending on the predictive score, and then adjusts for confounding using a pre-specified outcome model, for instance, using a simple linear regression of the outcome on both the explanatory variables and the surrogate variables for the confounders estimated using the factor model. We do not present them in this essay, but the predictive scores' definition and different outcome models are given by Wang and Blei (2019) (Section 4). Finally, although it aims to provide causal estimates, the procedure bases itself on few assumptions: the standard ones of *consistency* and *positivity*, along with the assumption that no confounder affects only one explanatory variable (*no unobserved single-cause confounders*).

The work of Wang and Blei (2019) sparked many discussions in this area of research and has been heavily commented on and critiqued for its ambitious claims. In particular, D'Amour (2019) challenged the identification conditions for the model and showed using counter-examples that general non-parametric identification is infeasible without additional assumptions. Ogburn et al. (2020) challenged the theory of the presented method and further showed that, without additional assumptions, the algorithm fails to control for multi-cause confounding. Grimmer et al. (2020) presented a thorough analysis of the approach and the theoretical results of Wang and Blei (2019), and showed using multiple case studies that it does not, in general, outperform a naive semi-parametric regression that does not take confounding into account. They highlight additionally that the assumptions of Wang and Blei (2019) may be difficult to satisfy in practice by showing that they require that each confounder affects infinitely many explanatory variables.

4 Confounding in multiple hypothesis testing

In this section, we now focus on adjusting for confounding in multiple hypothesis testing. To do so, we rely on the work of Wang et al. (2017), which brings together multiple confounding adjustment methods under a common framework and give theoretical guarantees for these methods under some assumptions.

We first illustrate how confounding can affect statistical analyses when doing multiple hypothesis testing using example E1. Once again, recall that we argued that ethnicity or geographical origin was a potential confounder when studying the genetic determinants of cardiovascular diseases. Suppose that the researcher is this time interested in finding potentially interesting genetic variations that are related to cardiovascular diseases. She wants to simultaneously test the significance of the association between each genetic variation in the data and some measure of whether the individuals have experienced some cardiovascular disease. However, as we argued, geographical origin, which is unobserved in this case, is correlated with both the genetic variations variables and the probability of having developed cardiovascular disease. This poses two main problems: first, because the confounder is correlated with multiple genetic variations, the test statistics are correlated with one another, and thus the tests cannot be assumed to be independent of each other anymore. Secondly, as the confounder is correlated with both the genetic variants and the variable measuring cardiovascular diseases, the tests statistics are also confounded.

In Section 3, we argued that under some assumptions, one could use principal components as surrogate variables to estimate the confounding factors, and thus adjust for confounding using spectral transformations. In this case, because we are switching the focus from regression to multiple hypothesis testing, this would amount to computing the singular value decomposition of the matrix Y (instead of the design matrix X Section 3) and applying a spectral transformation like PCA adjustment the same way as described earlier. However, one of the pitfalls of doing so is that the principal components whose contribution we decrease may be correlated with the primary effect of interest, in which case we would be removing the signal of interest from the data by trying to adjust for confounding.

In contrast, in the following section, we briefly present the method developed by Wang et al. (2017) which relies on factor models to obtain estimators of the underlying effect and valid asymptotic tests despite unmeasured confounding.

4.1 Modelling assumptions

Consider the linear confounding model M2. Recall that in this model, we suppose the data for n observations originates from a high-dimensional linear model where the outcomes $Y_1, ..., Y_q$ can be written as a linear combination of both the primary variable of interest X and the confounding variables $Z_1, ...Z_r$, as well as a noise matrix E. We also assume that we can write $Z = X\alpha + W$, where the confounding variables are correlated with the primary variable of interest X, so $\alpha \neq 0$, and where W is a random noise matrix independent of X and U. In this model, β describes the primary

effect of the observed covariate X on the outcomes Y that we wish to estimate and test. The model M2 can thus be written in the following form:

$$\begin{aligned} Z &= X\alpha + W, \\ Y &= X\beta + Z\Pi + U. \end{aligned} \tag{M2}$$

Furthermore, we assume in the following that the entries of X are distributed i.i.d. from a univariate distribution with mean zero and variance equal to 1 (but not necessarily a Gaussian). Additionally, we let the rows of W be normally distributed with mean zero and covariance matrix I_r, so $\forall i \in \{1, ..., n\} : W_i \sim \mathcal{N}(0, I_r)$); and the rows of U be normally distributed with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{q \times q}$ diagonal with entries $\sigma_1^2, ..., \sigma_q^2$, so $\forall i \in \{1, ..., n\} : U_i \sim \mathcal{N}(0, \Sigma))$. Finally, we assume that $\forall i, j \in \{1, ..., n\} : W_i \perp\!\!\!\perp X_j$ and $U_i \perp\!\!\!\perp (X_j, Z_j)$.

4.2 Outline of the two-step estimation procedure

Using these modelling assumptions, we introduce in this section the two-step inference procedure presented by Wang et al. (2017). In short, the method uses a factor model on a residual matrix, where X has been regressed out from each column of Y independently in the first step, and then uses the factor model estimates in the second step to remove the confounding effect from an initial (naive) estimate of β that does not take the confounding effect into account.

First step Consider first the expression $Y = X\beta + Z\Pi + U$ in M2. We begin by computing $\check{Y}$ a matrix of residuals by regressing out X on each column of Y independently. Using ordinary least squares, notice that by writing $Y^{(i)}$ and $\check{Y}^{(i)}$ for the i-th column of Y and $\check{Y}$ respectively, we have:

$$\check{Y} = (\check{Y}^{(1)}, ..., \check{Y}^{(q)}) \quad \text{such that} \quad \forall i \in \{1, ..., q\} : \check{Y}^{(i)} = Y^{(i)} - X(X^TX)^{-1}X^TY^{(i)}.$$

Thus, by writing $H_X = X(X^TX)^{-1}X^T$ for the projection matrix of X, we can rewrite the matrix $\check{Y}$ in the following form:

$$\check{Y} = Y - X(X^TX)^{-1}X^TY = (I_n - H_X)Y.$$

Here, $\check{Y}$ is the estimated counterpart of $Y - X\beta$ in the expression $Y = X\beta + Z\Pi + U$. By rewriting this expression as $Y - X\beta = Z\Pi + U$, we retrieve the form of a factor analysis model in which we can think of $Y - X\beta$ as the observed matrix, and where Z are the factors, Π is the factor loading matrix, and U is an error matrix. Building on this representation, the first step of the inference procedure consists of estimating the factor loading matrix Π using a factor analysis model on $\check{Y}$. We have briefly presented an estimating procedure for a factor model in Section 2.2.3. However, we consider a more general case here and will introduce how this is done in practice in an asymptotic setting where both $p, n \to \infty$ in the following section.

Second step Notice now that, analogously to how we switched back and forth between the linear confounding model and the perturbed linear model in Section 3, we can rewrite $Y = X\beta + Z\Pi + U$ by replacing Z with its expression from M2, so that:

$$Y = X\beta + (X\alpha + W)\Pi + U = X(\beta + \alpha\Pi) + (U + W\Pi).$$

Furthermore, since U and W are both independent of X, the term $U + W\Pi$ can be viewed as an error term in this linear model. Thus, when regressing out X in the first step of the procedure, by the usual properties of ordinary least squares, we obtain an unbiased estimate of τ, where:

$$\tau = \beta + \alpha\Pi.$$

However, recall that we are in this essay truly interested in estimating the unconfounded effect β, and not necessarily the confounded one τ. To obtain an estimate for β, notice first that the expression $\tau = \beta + \alpha\Pi$ can be seen as a linear model, in which we have already derived an estimate for the factor loading matrix Π in the first step, and where we can also obtain an unbiased estimate for τ by looking at the coefficient of the regression of X on each column of Y in the first step. Then, the second step of the inference procedure builds on this representation as a linear model to estimate α first, and then use the relationship $\beta = \tau - \alpha\Pi$ to estimate β.

4.3 Identifiability of the parameters

The inference procedure presented in Section 4.2 provides an intuitive way of using factor models for inference under unmeasured confounding. However, with no additional assumption, the parameters β and α are not identifiable in this model. In this section, we present some assumptions under which they become identifiable. For this, recall that the parameters in the model are α, β, Π and the covariance matrix Σ, and let Θ denote their parameter space.

Following Wang et al. (2017) and Sun et al. (2012), we start by introducing an orthogonal transformation matrix Q that will allow us to separate the inference of Γ and Σ from that of α and β, and investigate their identifiability. Let Q be the Householder transformation matrix so that we have:

$$Q = I_n - 2vv^T \quad \text{where} \quad v = \frac{1}{\|X - e_1\|_2}(X - e_1) \quad \text{and} \quad e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^n.$$

By definition of the Householder matrix Q, we have $Q^T X = \|X\|_2^2 e_1$ and we can thus rewrite the equations from M2 as follows:

$$\tilde{Z} := Q^T Z = Q^T X\alpha + Q^T W = \|X\|_2^2 e_1\alpha + Q^T W,$$
$$\tilde{Y} := Q^T Y = Q^T X\beta + Q^T Z\Pi + Q^T U = \|X\|_2^2 e_1\beta + \tilde{Z}\Pi + Q^T U.$$

This representation allows us to separate the first row of $\tilde{Y}$ from the rest and simplify the problem. Writing $\tilde{W}$ and $\tilde{U}$ for $Q^T W$ and $Q^T U$ respectively, and indexing by -1 for a matrix composed of all rows but the first, we have:

$$\begin{aligned} \tilde{Z}_1 &= \|X\|_2^2 \alpha + \tilde{W}_1, \\ \tilde{Z}_{-1} &= \tilde{W}_{-1}, \\ \tilde{Y}_1 &= \|X\|_2^2 \beta + \tilde{Z}_1 \Pi + \tilde{U}_1, \\ \tilde{Y}_{-1} &= \tilde{Z}_{-1} \Pi + \tilde{U}_{-1}. \end{aligned} \tag{M5}$$

Using this representation allows us to separate the inference problem of α and β from that of Π and Σ. To see this, notice first that α and β are only present in the expression $\tilde{Y}_1$ (but not in $\tilde{Y}_{-1}$), so we can use the first column of $\tilde{Y}$ to do inference for the parameters α and β.

Then, since the rows of W are distributed i.i.d, because Q is unitary by definition, the rows of $\tilde{W} = Q^T W$ will have the same distribution as the rows of W. The same applies for $\tilde{U}$, and thus the rows of $\tilde{U}$ also have a multivariate normal distribution with mean zero and covariance matrix Σ. Furthermore, as we assumed that the rows of U (and hence of $\tilde{U}$) were distributed i.i.d. from a multivariate normal distribution, we have that $\tilde{W}_1 \perp\!\!\!\perp \tilde{W}_{-1}$ and thus $\tilde{Y}_1 \perp\!\!\!\perp \tilde{Y}_{-1} \mid X$. Hence, we can treat the inference of Π, and α and β as two different (successive) problems.

Finally, it follows from this decomposition that $\tilde{Y}_1 = \|X\|_2^2(\beta + \alpha\Pi) + \tilde{W}_1\Pi + \tilde{U}_1$ and hence we can write:

$$\begin{aligned} \tilde{Y}_1 = \|X\|_2^2(\beta + \alpha\Pi) + \tilde{W}_1\Pi + \tilde{U}_1 \quad &\Rightarrow \quad \tilde{Y}_1 \sim \mathcal{N}(\|X\|_2^2(\beta + \alpha\Pi), \Sigma + \Pi^T\Pi), \\ \tilde{Y}_{-1} = \tilde{W}_{-1}\Pi + \tilde{U}_{-1} \quad &\Rightarrow \quad \forall i \in \{1, ..., n-1\} : \tilde{Y}_{-1,i} \sim \mathcal{N}(0, \Sigma + \Pi^T\Pi). \end{aligned}$$

Identifiability of Π Using this decomposition, the expression $\tilde{Y}_{-1} = \tilde{Z}_{-1}\Pi + \tilde{U}_{-1}$ is exactly that of a factor analysis model in which $\tilde{Y}_{-1}$ is the observed matrix, $\tilde{Z}_{-1}$ are the factors, and Π is the factor loading matrix. Notice first that in the setting where we are mostly interested in β, it does not matter if Π is identified only up to a rotation. In fact, using the re-parametrisation $\Pi \leftarrow U\Pi$ and $\alpha \leftarrow aU^T$ where $U \in \mathbb{R}^{r \times r}$ is an orthogonal matrix does not impact the value of β in the expression of $\tilde{Y}_1$.

Building on this, Wang et al. (2017) use the following result from the literature on factor analysis to ensure the identifiability of Π. Lemma 4.1 is only stated here, but the proof of the more general result is given in full by Wang et al. (2017) (Lemma 2.1) or by Anderson et al. (1956) (Theorem 5.1).

Lemma 4.1. *A sufficient condition for Π and Σ to be identifiable in model M5 up to multiplication on the left by an orthogonal matrix is that if any column of Π is deleted, there remain two disjoint submatrices of rank r.*

We denote by $\Theta_0 \subset \Theta$ the parameter space where Π satisfies the conditions of Lemma 4.1, i.e. where removing a column from Π always yields two disjoint submatrices of rank r. Notice that as $\Pi \in \mathbb{R}^{r\times q}$, this can be true only if $q > 2r + 1$.

Identifiability of α and β The parameters α and β cannot generally be identified without additional assumptions. Remark that $\tilde{Y}_1$ is a vector in $\mathbb{R}^{1\times q}$, but that $\alpha \in \mathbb{R}^{1\times r}$ and $\beta \in \mathbb{R}^{1\times q}$, so we have in total $q + r$ parameters to estimate.

One way to overcome this identification problem is to look at the cases where β contains at least r entries equal to zero and to assume one of these cases to be true when making inferences. Wang et al. (2017) consider two sufficient conditions: first, the case where the researcher has access to at least r controls for which he knows that X does not affect the outcome Y; and secondly, the case where the vector of coefficients β is assumed to be sparse in the sense that at least half of the true effects are null. Both of these cases are typical in genetic studies, for example, when using genotyping microarrays where the researcher often has access to a negative control set to assess the quality of the data; or in GWAS where she may expect that only a small number of genes are related to the studied medical condition. These two restrictions of the parameter space are as follows.

Negative control $\Theta_1 = \{(\alpha, \beta, \Pi, \Sigma) : \beta_{\mathcal{C}}^T = 0, \text{rank}(\Pi_{\mathcal{C}}^T) = r\}$ for a known control set $\mathcal{C}$ such that $|\mathcal{C}| \geq r$.

Sparsity $\Theta_2(s) = \{(\alpha, \beta, \Pi, \Sigma) : \|\beta^T\|_0 \leq \lfloor (q - s)/2 \rfloor, \text{rank}(\Pi_{\mathcal{C}}^T) = r, \forall \mathcal{C} \subset \{1, ..., q\}, |\mathcal{C}| = s\}$ for some $r \leq s \leq q$.

Following Wang et al. (2017), we show that α and β are identifiable under both of these restrictions, given that $(\alpha, \beta, \Pi, \Sigma) \in \Theta_0$, that is when Π is identifiable.

Lemma 4.2. *If $\Theta = \Theta_0 \cap \Theta_1$ or $\Theta = \Theta_0 \cap \Theta_2(s)$ for some $r \leq s \leq q$, the parameters $(\alpha, \beta, \Pi, \Sigma)$ in model M5 are identifiable.*

Proof. Since $\Theta \subset \Theta_0$, we know that Π and Σ are identifiable. Recall that the parameters $(\alpha, \beta, \Pi, \Sigma)$ are identifiable if the mapping between these parameters and the distribution in M5 is one-to-one. Consider two combinations of parameters $(\alpha^{(1)}, \beta^{(1)}, \Pi, \Sigma)$ and $(\alpha^{(2)}, \beta^{(2)}, \Pi, \Sigma)$ both in Θ, where $\beta^{(1)} + \alpha^{(1)}\Pi = \beta^{(2)} + \alpha^{(2)}\Pi$ so that $\tilde{Y}$ in M5 is equal for these two parameter sets. Let $\mathcal{C}$ denote the set of parameters for which $\beta_{\mathcal{C}}^{(1)} = \beta_{\mathcal{C}}^{(2)} = 0$. We distinguish the two cases highlighted above:

- if $\Theta = \Theta_0 \cap \Theta_1$, we have $\beta^{(1)} + \alpha^{(1)}\Pi = \beta^{(2)} + \alpha^{(2)}\Pi \Rightarrow \Pi_{\mathcal{C}}^T \alpha^{(1)T} = \Pi_{\mathcal{C}}^T \alpha^{(2)T}$. Recall that for any matrix $A \in \mathbb{R}^{m\times k}$ and two vectors $X, Y \in \mathbb{R}^k$, we have $AX = AY \Rightarrow X = Y$ if A has full column rank and $m > k$. Here, we know that $|\mathcal{C}| \geq r$ and $\text{rank}(\Pi_{\mathcal{C}}^T) = r$ so $\Pi_{\mathcal{C}}^T \in \mathbb{R}^{|\mathcal{C}|\times r}$ has full column rank, and hence we have $\alpha^{(1)} = \alpha^{(2)}$.

- if $\Theta = \Theta_0 \cap \Theta_2(s)$, we can also show that $\Pi_{\mathcal{C}}^T \alpha^{(1)T} = \Pi_{\mathcal{C}}^T \alpha^{(2)T}$ where $|\mathcal{C}| = s$ and where $\text{rank}(\Pi_{\mathcal{C}}^T) = r$ with $r < s$, so $\Pi_{\mathcal{C}}^T \in \mathbb{R}^{s\times r}$ has full column rank, and hence we have $\alpha^{(1)} = \alpha^{(2)}$.

Hence, in both of these cases $\alpha^{(1)} = \alpha^{(2)}$, which implies $\beta^{(1)} = \beta^{(2)}$. □

4.4 Inference and hypothesis testing

Building on these results on identifiability, Wang et al. (2017) propose a general method for estimating β in both scenarios (with negative controls or sparsity). In this section, we briefly present and state some of their results.

To adhere more closely to the notations from Wang et al. (2017) and for clarity in the presented results, we write from now on $\Gamma = \Pi^T$, so that M5 becomes:

$$\begin{aligned}
\tilde{Z}_1 &= \|X\|_2^2\alpha + \tilde{W}_1, \\
\tilde{Z}_{-1} &= \tilde{W}_{-1}, \\
\tilde{Y}_1 &= \|X\|_2^2\beta + \tilde{Z}_1\Gamma^T + \tilde{U}_1, \\
\tilde{Y}_{-1} &= \tilde{Z}_{-1}\Gamma^T + \tilde{U}_{-1}.
\end{aligned}$$

Inference for Π / Γ and Σ We have seen in Section 4.2 and Section 4.3 that Π (or Γ) and Σ could be estimated using a factor model. In fact, Wang et al. (2017) (Section 5.3.1) show that the more intuitive estimator from Section 4.2 (where the factor model uses the residual matrix $\check{Y}$ as the observed matrix) yields the same results as the estimator from the factor model in model M5 which uses $\tilde{Y}_{-1}$ as the observed matrix.

However, standard methods for exploratory factor analysis do not provide consistent estimates of Γ and Σ in the asymptotic setting where both $n \to \infty$ and $p \to \infty$. To overcome this problem, Wang et al. (2017) use a result derived by Bai et al. (2012) which provides theoretical guarantees of convergence of the estimated factor loadings and the diagonal covariance Σ using quasi-log-likelihood. We only state the result here, but the full form of the quasi-log-likelihood, along with further detail and the proof, are given in full by Wang et al. (2017) (Section 3.1).

Lemma 4.3. *Let Γ and Σ be such that:*

(i) the noise matrix U has rows normally distributed with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{q\times q}$ diagonal with entries $\sigma_1^2, ..., \sigma_q^2$,

(ii) there exists a $D > 0$ such that $\|\Gamma_j\|_2 \leq D$ and $\forall i \in \{1, ..., q\} : D^{-2} \leq \sigma_i^2 \leq D^2$ and $\forall i \in \{1, ..., q\} : D^{-2} \leq \hat{\sigma}_i^2 \leq D^2$ where $\hat{\sigma}_1^2, ..., \hat{\sigma}_q^2$ are the estimated variances,

(iii) the limits $\lim_{q\to\infty} q^{-1}\Gamma^T\Sigma^{-1}\Gamma$ and $\lim_{q\to\infty} \sum_{j=1}^q \sigma_j^{-4}(\Gamma_j \otimes \Gamma_j)(\Gamma_j^T \otimes \Gamma_j^T)$ exist and are positive semi-definite matrices.

Let also $\Gamma^{(0)} = \Gamma R$ where $RR^T = (n-1)^{-1}\tilde{Z}_{-1}^T\tilde{Z}_{-1}$. Then, the maximisers $(\hat{\Gamma}, \hat{\Sigma})$ of the quasi-log-likelihood defined in Wang et al. (2017) (Section 3.1) satisfy:

$$\sqrt{n}(\hat{\Gamma}_j - \Gamma_j^{(0)}) \xrightarrow{d} \mathcal{N}(0, \sigma_j^2 I_r) \quad \text{and} \quad \sqrt{n}(\hat{\sigma_j^2} - \sigma_j^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma_j^4).$$

Hence, Γ (or Π per the notations from past sections) and Σ can be estimates from the data. We will use these estimates when estimating α and β.

Inference for α and β Following Wang et al. (2017), we present in this section a procedure for estimating α and β using the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ derived above. We only focus on the negative control scenario in the following.

Recall that α and β can be estimated using $\tilde{Y}_1 = \|X\|_2^2\beta + \tilde{Z}_1\Gamma^T + \tilde{U}_1$ from the model M5, which we can rewrite as:

$$\tilde{Y}_1/\|X\|_2^2 = (\beta + \alpha\Gamma^T) + (\tilde{W}_1/\|X\|_2^2)\Gamma^T + \tilde{U}_1/\|X\|_2^2.$$

To use the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ derived above, notice that we can rewrite this expression by taking $\Gamma^{(0)} = \Gamma R$ and $\alpha^{(0)T} = R^{-1}(\alpha^T + \tilde{W}_1^T/\|X\|_2^2)$ where R only depends on $\tilde{Y}_{-1}$ and is thus independent of $\tilde{Y}_1$, so that, by transposing, we can write:

$$\tilde{Y}_1^T/\|X\|_2^2 = \beta^T + \Gamma^{(0)}\alpha^{(0)T} + \tilde{U}_1/\|X\|_2^2.$$

Negative control Suppose that we are in the scenario where we have a set $\mathcal{C}$ of at least r controls for which we know that $\beta_{\mathcal{C}} = 0$. Then, we can rewrite the expression above as follows:

$$\tilde{Y}_{1,\mathcal{C}}^T/\|X\|_2^2 = \Gamma_{\mathcal{C}}^{(0)}\alpha^{(0)T} + \tilde{U}_{1,\mathcal{C}}/\|X\|_2^2,$$
$$\tilde{Y}_{1,\mathcal{C}}^T/\|X\|_2^2 = \beta_{\mathcal{C}}^T + \Gamma_{-\mathcal{C}}^{(0)}\alpha^{(0)T} + \tilde{U}_{1,-\mathcal{C}}/\|X\|_2^2.$$

Building on this representation, Wang et al. (2017) propose the following negative control estimator which successively estimates $\alpha^{(0)}$ using generalised least squares, and where the estimator for β follows from the linear relationship between α, β and Γ from model M5:

$$\hat{\alpha}^{NC} = (\hat{\Gamma}_{\mathcal{C}}^T\hat{\Sigma}_{\mathcal{C}}^{-1}\hat{\Gamma}_{\mathcal{C}})^{-1}\hat{\Gamma}_{\mathcal{C}}^T\hat{\Sigma}_{\mathcal{C}}^{-1}\tilde{Y}_{1,\mathcal{C}}^T/\|X\|_2^2,$$
$$\hat{\beta}_{-\mathcal{C}}^{NC} = \tilde{Y}_{1,-\mathcal{C}}^T/\|X\|_2^2 - \hat{\Gamma}_{-\mathcal{C}}\hat{\alpha}^{NC}.$$

Then, letting $\Sigma_{\mathcal{C}}$ represent the covariance matrix of the variables in $\mathcal{C}$, it can be shown that $\hat{\beta}_{-\mathcal{C}}^{NC}$ converges to $\beta_{-\mathcal{C}}^{NC}$ under additional assumptions.

Theorem 4.4. *Under the assumptions (i), (ii) and (iii) from Lemma 4.3, assume additionally that:*

(iv) $\lim_{p\to\infty} |\mathcal{C}|^{-1}\Gamma_{\mathcal{C}}^T\Sigma_{\mathcal{C}}^{-1}\Gamma_{\mathcal{C}}$ *exists and is positive definite.*

Then, if $n \to \infty$ *and* $p \to \infty$ *where* $(\log p)^2/n \to \infty$*, for any fixed index set with finite cardinality* $\mathcal{S}$ *such that* $\mathcal{S} \cap \mathcal{C} = \varnothing$*, we have:*

$$\sqrt{n}(\hat{\beta}_{\mathcal{S}}^{NC} - \beta_{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, (1 + \|\alpha\|_2^2)(\Sigma_{\mathcal{S}} + \Delta_{\mathcal{S}})),$$

where $\Delta_{\mathcal{S}} = \Gamma_{\mathcal{S}}(\Gamma_{\mathcal{C}}^T\Sigma_{\mathcal{C}}^{-1}\Gamma_{\mathcal{C}})^{-1}\Gamma_{\mathcal{S}}^T$*. In particular, if* $|\mathcal{C}| \to \infty$*, then by (iv) the largest eigenvalue of* $\Gamma_{\mathcal{C}}^T\Sigma_{\mathcal{C}}^{-1}\Gamma_{\mathcal{C}} \to \infty$ *and then the maximum entry of* $\Delta_{\mathcal{S}}$ *goes to zero, thus:*

$$\sqrt{n}(\hat{\beta}_{\mathcal{S}}^{NC} - \beta_{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, (1 + \|\alpha\|_2^2)\Sigma_{\mathcal{S}}).$$

Notice that the asymptotic variance in the case where $|\mathcal{C}| \to \infty$ is the same as the asymptotic variance that one would obtain using an ordinary least squares regression where the confounders are observed. In fact, when regressing out both X and Z on the j-th column of Y in model M2, we can write the model as:

$$Y^{(j)} = X\beta_j + Z\Gamma_j^T + E^{(j)} \quad \text{where} \quad E^{(j)} \sim \mathcal{N}(0, \sigma_j I_r).$$

Thus, the asymptotic variance of the OLS estimator $(\hat{\beta}_j^{\text{OLS}}, \hat{\Gamma}_j^{\text{OLS}})$ is given by the usual formula $\sigma_j^2 \text{Var}([X_i, Z_i])^{-1}/n$, from which we can obtain under our distributional assumptions that $\text{Var}(\hat{\beta}_j^{\text{OLS}}) = \sigma_j^2(1 + \|\alpha\|_2^2)/n$ and $\forall j \neq k : \text{Cov}(\hat{\beta}_j^{\text{OLS}}, \hat{\beta}_k^{\text{OLS}}) = 0$. Thus $\text{Var}(\hat{\beta}^{\text{OLS}}) = (1/n)(1 + \|\alpha\|_2^2)\Sigma$, which we also find in Theorem 4.4.

Apart from the assumptions which guarantee the asymptotic convergence of the estimates $(\hat{\Gamma}, \hat{\Sigma})$, notice that the conditions under which this result holds are first, that the effect of the confounders is strong enough, including on the controls (which is related to assumptions (iii) and (iv)); and second, that the number of negative controls increases with p to infinity, i.e. such that $|\mathcal{C}| \to \infty$.

Hypothesis testing Suppose now that $|\mathcal{C}| \to \infty$ (so we are in the particular case of Theorem 4.4). Then, using the estimator $\hat{\beta}_j^{NC}$, we can construct a test statistic for the asymptotic test $H_{0,j} : \beta_j = 0$ against $H_{1,j} : \beta_j \neq 0$ in the usual way, where:

$$t_j = \frac{\|X\|_2 \hat{\beta}_j^{NC}}{\hat{\sigma}_j \sqrt{1 + \|\hat{\alpha^{NC}}\|_2^2}}.$$

Then, the null hypothesis $H_{0,j}$ is rejected at level α when $|t_j| > \Phi^{-1}(1 - \alpha/2)$. Hence, under the assumptions given above, the effect of the covariate X on a single of the outcomes (i.e. on a column of Y) can be tested in the usual way, with the estimators $\hat{\alpha}^{NC}$ and $\hat{\beta}_j^{NC}$.

This does not imply, however, that the effect of multiple covariates can be tested using standard methods such as Bonferroni correction or the Benjamini-Hochberg procedure, because the tests statistics derived above are not mutually independent. Nonetheless, Wang et al. (2017) show that despite this, the overall type-I error and the family-wise error rate (FWER) can be controlled.

Theorem 4.5. *Let $\mathcal{N}_q = \{j \in \{1, ..., q\} \mid \beta_j = 0\}$. Then, under the assumptions of Theorem 4.4 and when $|\mathcal{C}| \to \infty$, as $n, p, |\mathcal{N}_q| \to \infty$, we have:*

$$\frac{1}{|\mathcal{N}_q|} \sum_{j \in \mathcal{N}_j} \mathbb{1}(|t_j| > \Phi^{-1}(1 - \alpha/2)) \xrightarrow{q} \alpha,$$

$$\limsup \mathbb{P}\left(\sum_{j \in \mathcal{N}_j} \mathbb{1}(|t_j| > \Phi^{-1}(1 - \alpha/(2q))) \right) \leq \alpha.$$

Finally, Wang et al. (2017) introduce a method to test for potential confounding based on this inference procedure. Notice first that when $\alpha = 0$ in model M2, the variables Z aren't confounders anymore, but only unobserved variables which affect the outcomes Y but not the primary variable of interest X. Building on this, one may want to test the null hypothesis $H_0 : \alpha = 0$, and to adjust for confounding if this null hypothesis is rejected. Following Wang et al. (2017), Theorem 4.6 introduces a test for the confounding effect of the latent factors based on the estimate $\hat{\alpha}^{NC}$.

Theorem 4.6. *Let the assumptions of Theorem 4.4 be given, and let $\hat{\alpha} = \hat{\alpha}^{NC}$. Then, when $|\mathcal{C}| \to \infty$ as $n, p, |\mathcal{N}_q| \to \infty$, under the null hypothesis $H_0 : \alpha = 0$, we have:*

$$n \cdot \hat{\alpha}^T \hat{\alpha} \xrightarrow{d} \chi^2_r,$$

where χ^2_r is the chi-square distribution with r degrees of freedom.

4.5 Case study: the impacts of tourism in Germany

In this final section, we continue to investigate example E2 on tourism in Germany and use the approach from Wang et al. (2017) to illustrate how one may adjust for potential confounding bias in practice. Here, we are more interested in the impact of tourists' visits on a city rather than in the determinants of tourism. We use the same dataset as in Section 3.1.4 with $n = 101$ cities, and $p = 206$ continuous covariates. Notice again that our relatively low sample size and covariates are close to the limit case for applying such methods in practice.

We consider the task of testing the associations between tourism and a set of covariates. As we are interested in doing multiple hypothesis testing, notice that we change the labels from Section 3.1.4 and consider this time $X \in \mathbb{R}^n$ to be the primary variable of interest (the number of nights spent in tourist accommodation per 1000 inhabitants and year), and $Y \in \mathbb{R}^{n \times p}$ the matrix of outcomes whose association with the primary variable of interest we wish to test. Another important difference with the approach taken in Section 3.1.4 is that, both before or after adjusting for confounding, we now test multiple associations between a single covariate and a single outcome, whereas we were previously regressing multiple covariates on a single outcome, thus controlling for the effect of other covariates when studying the association between a specific explanatory variable and the outcome. Hence, we are more interested here in the changes induced by tourism on German cities than in what drives tourism. Finally, notice that we allow ourselves to comment on the results obtained here, but acknowledge that different comments could be drawn from the same results.

Multiple hypothesis testing with no adjustment We first test the associations between the primary variable of interest X and the outcomes Y without adjusting for any confounding. To do so, we use an ordinary least squares linear regression of the primary variable (the number of nights spent in tourist accommodation per 1000 inhabitants and per year) on each outcome variable, and control the family-wise error rate (FWER) at a 5% level using the Bonferroni procedure. After controlling the family-wise error rate, tourism is significantly associated with 40 outcome variables.

We do not detail the complete list of outcome variables in this section, but the associated code files can be found in the GitHub repository for reproducibility. In short, we observe that tourism is associated with many outcome variables, including rather intuitive ones like the number of museum visitors per year or the average price for buying a house or an apartment. However, we argue that some other associations are less intuitive, like a positive association with the birth rate and a negative one with the death rate, and are potentially due to confounding factors.

Testing for potential confounding Building on the intuition that the associations may hide confounding factors, we use the approach from Wang et al. (2017) to adjust for confounding. Because we do not have access to negative controls within the dataset, we use the sparsity scenario of the estimating procedure, which relies on robust regression and which can be found in full detail in Wang et al. (2017) (Section 3). Notice that this scenario applies when at least half of the true effects are null, which is only assumed here and cannot be verified, but only argued to be more or less likely to hold given the variables in the dataset.

First, using Theorem 4.6, we test for the presence of confounders in this study by testing the null hypothesis that the estimated factors are not correlated with the primary variable of interest X. For this dataset, the $p-$value of this test is 1.15×10^{-29} which strongly indicates the presence of confounders.

Multiple hypothesis testing with confounding adjustment Building on this, we try to adjust for confounding using the method of Wang et al. (2017). To do so, we first estimate the number of confounders using bi-cross validation, which is also used by Wang et al. (2017) and was introduced by Owen et al. (2016). In this setting, bi-cross validation suggests adjusting for $r = 11$ factors in the estimating procedure.

The resulting estimates of the confounding adjustment method $\hat{\beta}_{\texttt{CODE}}$, which model the association between tourism and the outcome variables, along with the associated $p-$values controlled at 5% FWER are presented in Table 3. Notice that only the significant associations are displayed, and that all outcome variables and the primary variable of interest have been standardised to have mean zero and unit variance.

We can first notice that the number of significant relationships is smaller than in the case with no adjustment for confounding. The remaining associations can be grouped in roughly six different themes: number of museum visitors per year, population levels, geographical origin and nationality, type of households, economic activity, and transport. Although being very significant, most coefficients are close to zero except for: the number of museum visitors per year; nationality and geographical origin (foreign-born inhabitants and foreigners); employment in financial and insurance activities; and transport (people commuting into the city).

Although these results give insights on the underlying relationships, notice importantly that, as with the approach from Section 3.1.4, the model may be misspecified, and that some of its underlying assumptions may not be verified. Hence these associations may not reflect true relationships, and should be interpreted with caution.

CODE	Outcome variables	$\hat{\beta}_{\texttt{CODE}}$	$p-$value
CR1007V	Number of museum visitors (per year)	0.1166	1.04×10^{-2}
DE1001V	Population on the 1st of January, total	0.0045	1.00×10^{-16}
DE1040V	Population on the 1st of January, 0-4 years, total	0.0242	1.00×10^{-16}
DE1074V	Population on the 1st of January, 5-9 years, total	0.0169	1.00×10^{-16}
DE1077V	Population on the 1st of January, 10-14 years, total	0.0079	1.00×10^{-16}
DE1058V	Population on the 1st of January, 25-34 years, total	0.0112	1.33×10^{-6}
DE1061V	Population on the 1st of January, 35-44 years, total	0.0090	5.65×10^{-3}
DE2009V	Foreign-born as a proportion of population	0.0810	1.84×10^{-3}
DE2012V	Foreigners as a proportion of population	0.0430	3.07×10^{-4}
DE3017V	Population living in private households	0.0048	1.00×10^{-16}
EC1001V	Economically active population, total	0.0019	1.31×10^{-11}
EC1174V	Economically active population, 20-64, total	0.0018	1.00×10^{-16}
EC1145V	Economically active population 55-64, total	-0.0093	1.00×10^{-16}
EC1177V	Persons employed, 20-64, total	0.0019	1.00×10^{-16}
EC1180V	Persons employed, 55-64, total	-0.0097	1.00×10^{-16}
EC2034V	Employment in financial and insurance activities	0.1489	2.43×10^{-2}
SA2007V	Number of live births per year	0.0250	1.00×10^{-16}
TT1064V	People commuting into the city	0.1021	1.37×10^{-2}

Table 3: Significant relationships (and respective coefficients and $p-$values) when testing associations between the primary variable of interest and the outcomes after adjusting for confounding and controlling at 5% FWER using the Bonferroni procedure

In particular, it is perhaps preferable to consider these results as a first step which would consist of identifying potentially interesting associations, before studying them further using different statistical methods, in the same way as multiple hypothesis testing is usually done to select candidate genes in genome-wide association studies. In fact, we can first emphasise that we observe numerous associations whose corresponding relationships are not intuitive, and would require further research, such as the negative association with the employment level of older populations, the population living in private households, or the number of live births.

Nonetheless, we can also see that we identify an association between tourism and the number of museum visitors per year again. As in Section 3.1.4, we argue that this association is relatively intuitive, although it would be difficult to comment on the direction of a potential causal link. Finally, tourism appears again to be positively associated with the number of foreigners and foreign-born inhabitants, although this time, the underlying coefficient estimate is positive.

5 Conclusion

High-dimensional data present new challenges and new opportunities to deal with unmeasured confounding. Whether the researcher wants to draw causal inferences from the data or, perhaps more realistically in practical cases, to simply study the associations between covariates and outcomes, having a large number of observed variables at hand not only strengthens the plausibility of unconfoundedness but also gives way to new estimating procedures that take potential confounding into account.

In this essay, we explored a few of these estimating procedures and their theoretical properties. In short, the approaches we presented rely on the idea that when the confounders affect many of the observed covariates, some information about them can be retrieved from the observed data. This information can then be used in the estimating procedures to adjust for the confounding bias. Following the literature on the subject, we showed that under the proper set of assumptions, the underlying effect of the explanatory variable(s) on the outcome(s) could be estimated with the same asymptotic properties as in the case where the confounders are observed.

We focused primarily on the case where the data originates from a linear confounding model. In this setting, we placed ourselves in both the case of high-dimensional regression and multiple hypothesis testing, and presented two confounding adjustment methods based on applying spectral transformations for the first, and using factor models for the second. In a non-linear setting, we briefly presented a novel and much-commented approach based on probabilistic factor models. These approaches rely all on strong assumptions about the data generating mechanisms and about the properties of the confounding effect. Nonetheless, they apply well to many practical settings like the discovery of genetic associations of diseases.

Finally, in a causal framework, for instance, when assuming that the linear models presented in this essay are structural equations models, these statistical approaches provide researchers with ways to estimate causal effects even when the *no unmeasured confounders* is not satisfied, provided that the underlying assumptions hold. However, intuitively, high-dimensional data may render the *positivity*, or *overlap* assumption more difficult to satisfy. Building on this observation, D'Amour et al. (2021) explore in detail the impact this can have on causal effect estimates in observational studies.

A Appendix

All empirical simulations, along with data for the case study can be found in the dedicated GitHub repository: `https://github.com/bglbrt/UCHDD/`.

We detail the structure of this repository in the following table:

File	Contents
`sim-fig4a.r`	Simulation code for Figure 4a (data generation and plotting)
`sim-fig4b.r`	Simulation code for Figure 4b (data generation and plotting)
`sim-fig5a.r`	Simulation code for Figure 5a (data generation and plotting)
`sim-fig5b.r`	Simulation code for Figure 5b (data generation and plotting)
`sim-fig6a.r`	Simulation code for Figure 6a (data generation and plotting)
`sim-fig6b.r`	Simulation code for Figure 6b (data generation and plotting)
`case-study1.r`	Case study code for Section 3.1.4 (graphs and tables)
`case-study2.r`	Case study code for Section 4.5 (tables)
`data`	
`data.csv`	Dataset for the case study
`var.csv`	List of variables and their respective codes
`preprocessing.py`	Preprocessing file to create the `data.csv` dataset
`raw_data`	
`urb_cecfi.tsv`	Original data file from `https://ec.europa.eu/eurostat/`
⋮	
`urb_geo.csv`	Original data file from `https://ec.europa.eu/eurostat/`

List of Figures

List of Tables

References

Anderson, T. W., Rubin, H., et al. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150.

Bai, J., Li, K., et al. (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics*, 40(1):436–465.

Ćevid, D., Bühlmann, P., and Meinshausen, N. (2018). Spectral deconfounding via perturbed sparse linear models. *arXiv preprint arXiv:1811.05352*.

D'Amour, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Grimmer, J., Knox, D., and Stewart, B. M. (2020). Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv preprint arXiv:2007.12702*.

Guo, Z., Ćevid, D., and Bühlmann, P. (2020). Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. *arXiv preprint arXiv:2004.03758*.

Johnstone, I. M. and Paul, D. (2018). Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292.

Miao, W., Hu, W., Ogburn, E. L., and Zhou, X. (2020). Identifying effects of multiple treatments in the presence of unmeasured confounding. *arXiv preprint arXiv:2011.04504*.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101.

Ogburn, E. L., Shpitser, I., and Tchetgen, E. J. T. (2020). Counterexamples to" the blessings of multiple causes" by wang and blei. *arXiv preprint arXiv:2001.06555*.

Owen, A. B., Wang, J., et al. (2016). Bi-cross-validation for factor analysis. *Statistical Science*, 31(1):119–139.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Sun, Y., Zhang, N. R., and Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, pages 1664–1688.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Van de Geer, S. (2007). The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5):1863.

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