University of California
Santa Barbara

Sensitivity Analysis for Causal Inference with Unobserved Confounding

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Jiajing Zheng

Committee in charge:

Professor Alexander Franks, Chair
Professor Sang-Yun Oh
Professor Yu-Xiang Wang

September 2021

The Dissertation of Jiajing Zheng is approved.

Professor Sang-Yun Oh

Professor Yu-Xiang Wang

Professor Alexander Franks, Committee Chair

September 2021

Sensitivity Analysis for Causal Inference with Unobserved Confounding

by

Jiajing Zheng

Acknowledgements

Words cannot begin to express my gratitude to my advisor, Professor Alex Franks, who has given me a great deal of encouragement and immense support along the way on this journey. I couldn't have accomplished this dissertation without all your assistance and dedicated involvement in every step throughout the process. Your expertise is invaluable in formulating the research questions and methodology, and your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would also like to show gratitude to my other thesis committee members, Professor Sang-Yun Oh and Yu-Xiang Wang for your constructive advice and general guidance.

I would also acknowledge the entire Department of Statistics and Applied Probability for their dedication and intelligence. In particular, Professor Rao Jammalamadaka, Andrew Carter, Yuedong Wang, Wendy Meiring, Michael Ludkovski and John Hsu, for their valuable guidance at the early stage of my studies. You provided me with the statistical basis that I needed for the following studies. And, of course, thanks should also go to the department staff for their kindness and strong backup.

I would also like to thank the entire lab for bringing up lots of interesting discussions and making many of my Fridays more colorful.

Finally, I would like to thank my friends, parents and beloved one, I could not have done this without all your company and support.

Curriculum Vitæ

Jiajing Zheng

Education

2021	Ph.D. in Statistics and Applied Probability (Expected), University of California, Santa Barbara.
2016	B.S. in Statistics, Wuhan University.

Publications

- Jiajing Zheng, Alexander D'Amour, and Alexander Franks. "Copula-based Sensitivity Analysis for Multi-Treatment Causal Inference with Unobserved Confounding." arXiv preprint arXiv:2102.09412 (2021)
- Alexander Franks, and Jiajing Zheng. "Bayesian Partial Identification for Muli-Treatment Inference with Unobserved Confounding." ICML, 2021.

Skills

- **Statistical**: Causal Inference, Sensitivity Analysis, Observational Studies, Multivariate Analysis, Predictive Modeling, Machine Learning, Regression Analysis, Bayesian Inference, Data Visualization, Analytical Thinking
- **Programming/Software**: R, SQL, LaTeX, SAS, Python, MATLAB, Excel, Word, PowerPoint

Work Experience

- Graduate Researcher / Teaching Assistant — UCSB, 2016-2021
- Data Science Intern — PayPal, 2020 Summer

Abstract

Sensitivity Analysis for Causal Inference with Unobserved Confounding

by

Jiajing Zheng

Many questions in social and biomedical sciences are causal in nature. For example, sociologists and policy-makers often want to know the effects of social programs on poverty and upward mobility; medical professionals are interested in how drugs impact the progression of disease. Unfortunately, estimating causal effects from non-experimental data is very difficult due to unobserved confounders, which can lead to spurious causal conclusions about the treatment's effect on the outcome. Sensitivity analysis, which explores how sensitive our causal conclusions are to potential unobserved confounding, can help us understand the potential impacts of confoundedness. However, existing sensitivity analyses are often at odds with modern machine learning (ML) tools for causal inference, which emphasize flexible models over interpretability. Besides, modern problems require new methods which account for the existence of multiple concurrent treatment variables and/or high dimensional outcomes. In this dissertation, we provide new tools that help improve communication and transparency about the robustness of analysis results to unmeasured confoundedness for important applications of observational causal inference, especially in high-dimensional settings. In chapter 1, we introduce the two most prevalent frameworks for causal inference studies, define the relevant quantities and notations, and discuss the importance of sensitivity analysis. In Chapters 2, we propose a sensitivity analysis method by reparametering latent confounder models, and in Chapter 3, we extend a sensitivity analysis method based on the Tukey's factorization to cases where treatments are ordinal variable with multiple

levels, where both methods clearly separate the identifiable part from the unidentifiable part. In Chapter 4 and 5, we focus on high-dimensional settings, respectively considering the multi-treatment and multi-outcome cases, where the multivariate correlation structure could provide additional information about unobserved confounders, but the causal effects are still not point identifiable in general and the high-dimensional variables would largely complicate the analysis. To address the issue, we present novel sensitivity analysis methods based on copula factorization, which can show how much is gained by leveraging latent structure in a given application while leave the observed data modeling untouched.

Contents

Curriculum Vitae **v**

Abstract **vi**

1 Introduction **1**
1.1 Causal inference in Potential Outcome Framework 2
1.2 Causal Inference in *do*-calculus framework 5
1.3 Sensitivity Analysis 7
1.4 Latent Confounder Models 10

2 Sensitivity Analysis with Reparameterized Latent Confounder Models **14**
2.1 Reparameterization 14
2.2 Calibration 17
2.3 Application: Analysis of NHANES data 19

3 Sensitivity Analysis via Tukey's Factorization with Ordinal Treatments **22**
3.1 Tukey's Factorization 22
3.2 Generalizing Tukey's Factorization 24
3.3 Logistic Selection with Exponential Family Models 26
3.4 Calibration 28
3.5 Simulation 30
3.6 Discussion 32

4 Copula-based Sensitivity Analysis with Multiple Treatments **34**
4.1 Introduction 35
4.2 Sensitivity Analysis via Copula Parameterizations with Multi-Treatment 37
4.3 Practical Sensitivity Analysis with the Gaussian Copula 41
4.4 The Geometry of Sensitivity in the Gaussian Copula Model 44
4.5 Calibration and Robustness 53
4.6 Simulation Studies 63
4.7 A Reanalysis of the Actor Case Study 71
4.8 Discussion 74

5 Copula-Based Sensitivity Analysis with Multiple Outcomes **77**
5.1 Introduction 78
5.2 Sensitivity Analysis via Copula Parameterizations with Multi-Outcome . 80
5.3 Sensitivity Analysis with Multiple Outcomes in the Gaussian Copula Model 82
5.4 Calibration 89
5.5 Analysis of Metabolomic Aging Clocks 95
5.6 Discussion 97

6 Discussion **99**

A Appendix for Chapter 4 **103**
A.1 Theory 103
A.2 Modeling Choice Details 114
A.3 Additional Results 119

B Appendix for Chapter 5 **122**
B.1 Theory 122
B.2 Additional Results 129

Bibliography **131**

Chapter 1

Introduction

Causal inference problems arises almost everywhere. For instance, in social science, sociologists and policy-makers often want to know the effects of social programs on poverty and upward mobility; in biomedical studies, medical professionals are interested in how drugs impact the progression of disease. Unfortunately, estimating causal effects from non-experimental data is especially difficult. The fundamental challenge is that unobserved confounders can bias our understanding of causal effects. A confounder is a variable that influences both the treatment and the outcome, which can lead to spurious causal conclusions about the treatment's effect on the outcome. For example, in studies on the effect of tobacco smoking on human health, alcohol consumption and unhealthy diet are potential confounders [1]. The detrimenntal effects of smoking might be overestimated if we did not control for these confounders. Sensitivity analysis, which explores the range of causal effects that are consistent with the observed data in the context of a given problem can help us understand the potential impacts of unmeasured confounders. In principle, a sensitivity analysis quantifies how results change under different assumptions about unobserved confounders without affecting the observed data model. Unfortunately, existing sensitivity analysis approaches perturb the observable predictions of the model

and degrade the quality of observed data predictions [2].

The remainder of this chapter proceeds as follows. In Section 1.1 and 1.2, we start by introducing of the two most common frameworks for causal inference: the potential outcome framework [3], developed by Donald B. Rubin, and the *do*-calculus framework [4], proposed by Judea Pearl. In Section 1.3, we introduce the formal setup for sensitivity analysis and highlight key problems about identifiability of sensitivity parameters. In Section 1.4, we describe one of the most common approaches to sensitivity analysis, called latent confounder models, and illustrate some difficulties with this approach by a simple example.

1.1 Causal inference in Potential Outcome Framework

Conventionally, in the potential outcome framework [5, 6], we use T to denote a binary treatment, with $T = 1$ indicating assignment to treatment and $T = 0$ indicating assignment to control, X to denote observed pretreatment variables, $Y_i(0)$ and $Y_i(1)$ to denote the outcomes that would be observed under $T = 0$ and $T = 1$ respectively. With the Stable Unit Treatment Value Assumption (SUTVA) (Assumption 1.1.1), we can write the observed outcome as $Y_i^{\text{obs}} = Y_i(1)T_i + Y_i(0)(1 - T_i)$.

Assumption 1.1.1 (SUTVA) *There are no hidden versions of the treatments and there is no interference between units (see [7]).*

Note that, in practice, we can only observe at most one of $Y_i(1)$ and $Y_i(0)$ for any individual. Therefore, it is impossible to measure causal effects at the individual level without additional strong assumptions. For this reason, researchers generally focus on estimating average treatment effects defined over the population. We define the Population Average

Treatment Effect (PATE) as

$$\text{PATE:} \quad E(Y(1) - Y(0)), \tag{1.1}$$

and relatively quantities, the Population Average Treatment Effect on the Treated/Control as

$$\text{PATT:} \quad E(Y(1) - Y(0) \mid T = 1), \tag{1.2}$$

$$\text{PATC:} \quad E(Y(1) - Y(0) \mid T = 0), \tag{1.3}$$

which correspond to the average difference in the pair of potential outcomes averaged over the treated and control respectively. Analogously, we define treatment effects in sub-populations stratified by observed covariates X, the Conditional Average Treatment Effect (CATE), the so-called Conditional Average Treatment Effect for the Treated (CATT) and Conditional Average Treatment Effect for the Control (CATT) respectively as

$$\text{CATE:} \quad \sum_{i=1}^{n} E(Y_i(1) - Y_i(0)|X_i), \tag{1.4}$$

$$\text{CATT:} \quad \sum_{i:T_i=1} E(Y_i(1) - Y_i(0)|X_i), \tag{1.5}$$

$$\text{CATT:} \quad \sum_{i:T_i=0} E(Y_i(1) - Y_i(0)|X_i). \tag{1.6}$$

For binary outcomes, researchers are usually more interested in the causal risk ratio (RR) between the the pair of potential outcomes,

$$\text{RR:} \quad E(Y(1))/E(Y(0)). \tag{1.7}$$

All estimands above are so-called "marginal contrast" estimands, meaning that they can

all be written as functions of the marginal complete-data outcome distributions of $Y(1)$ and $Y(0)$ [2]. We use $f(\cdot)$ and $F(\cdot)$ to respectively denote the probability density function and cumulative distribution function of random variables. For $t \in \{0, 1\}$, the complete-data density can be written as a mixture of the distribution of observed and missing outcomes:

$$\begin{aligned} f(Y(t) \mid X) =& f(T = t \mid X) f^{\text{obs}}(Y(t) \mid T = t, X) + \\ & f(T = 1 - t \mid X) f^{\text{mis}}(Y(t) \mid T = 1 - t), \end{aligned} \tag{1.8}$$

where $f^{\text{obs}}(Y(t) \mid T = t, X)$ is the observed outcome density, and $f^{\text{mis}}(Y(t) \mid T = 1 - t)$ is the missing outcome density, the only unidentifiable term. To identify the unobserved outcome densities, researchers assume strong ignorability of treatment assignment [8], which consists of the following two assumptions:

Assumption 1.1.2 (Ignorability in potential outcome framework)

$$[Y(0), Y(1)] \perp\!\!\!\perp T \mid X \tag{1.9}$$

Assumption 1.1.3 (Positivity in potential outcome framework)

$$0 \leq P(T = 1 \mid X) \leq 1 \tag{1.10}$$

Under the strong ignorability assumption of treatment assignment (Assumption 1.1.2 and 1.1.3), the estimands of interest can be identified by the observed data alone, since $f^{\text{obs}}(Y(t) \mid T = t, X) = f^{mis}(Y(t) \mid T = 1 - t, X)$. However, the strong ignorability assumption is untestable and unlikely to hold exactly in observational studies. We will come up with solutions that overcome this challenge in the later sections. Before that, in the next section, we would like to first introduce the other framework, the *do*-calculus

framework [4], that is also commonly used for causal inference problems.

1.2 Causal Inference in *do*-calculus framework

An Alternative to the potential outcome framework is the *do*-calculus framework, or the so-called causal graphical models, where probabilistic graphical models are used to encode assumptions about the data-generating process so that complex interrelationships between variables can be described concisely and implied properties can be read directly. Here, we let T denote the treatment variables, Y denote the outcome variables of interest, and t and y be realizations of the respective random variables, where both T and Y could be scalars or vectors, depending on the scenario we are considering. We let X denote any observed pre-treatment variables. In the *do*-calculus framework, $f(y \mid do(t))$ denotes the density of y in the population in which we have intervened to assign treatment level t to all units. In general, this is distinct from the observed outcome density, $f(y \mid t)$, which represents the density of the outcome in the subpopulation that received treatment t. These two densities are the same if and only if there are no confounders [9].

The goal of observational inference is to quantity the effects of different treatments by comparing the intervention distribution at different levels of treatment T [10, 11]. As before, we focus on *marginal contrast estimands* [2] here under arbitrary outcome and treatment distributions. We formalize the idea here again under the *do*-calculus framework. An estimand is a "marginal contrast" if it can be expressed as a function of the marginal distributions of the intervention outcomes, $\tau(E[v(y)|do(t_1)], E[v(y)|do(t_2)])$ for some functions v and τ. This includes the vast majority of commonly used estimands. For continuous outcomes, our primary estimand is the difference in the population average outcome for treatment $T = t_1$ and the population average outcome given treatment

$T = t_2$:

$$\text{PATE}_{t_1,t_2} := E(Y \mid do(t_1)) - E(Y \mid do(t_2)). \tag{1.11}$$

Here, $v(y) = y$ is the identity function and $\tau(a, b) = a - b$. When treatment is binary, the PATT and PATC can be expressed as following using the *do*-calculus framework:

$$\text{PATT:} \quad E(Y \mid do(t_1 = 1), T = 1) - E(Y \mid do(t_2 = 0), T = 1), \tag{1.12}$$

$$\text{PATC:} \quad E(Y \mid do(t_1 = 1), T = 0) - E(Y \mid do(t_2 = 0), T = 0). \tag{1.13}$$

We also consider the difference in the population average outcome receiving treatment t and the entire observed population average outcome, which we denote

$$\text{PATE}_{t,\bullet} := E(Y \mid do(t)) - E(Y), \tag{1.14}$$

where $E(Y) = \int E(Y \mid t) f(t) dt$ and $\text{PATE}_{t_1,t_2} = \text{PATE}_{t_1,\bullet} - \text{PATE}_{t_2,\bullet}$. The conditional average treatment effects are defined analogously as $\text{CATE}_{t_1,t_2|x} := E(Y \mid do(t_1), x) - E(Y \mid do(t_2), x)$ and $\text{CATE}_{t,\bullet|x} := E(Y \mid do(t), x) - E(Y \mid x)$.

For binary outcomes, our primary estimand is the causal risk ratio between treatments t_1 and t_2

$$RR_{t_1,t_2} := P(Y = 1|do(t_1))/P(Y = 1|do(t_2)). \tag{1.15}$$

where $RR_{t,\bullet}$ is defined analogously to (1.14), as $P(Y = 1 \mid do(t))/P(Y = 1)$, so that we can express $RR_{t_1,t_2} = RR_{t_1,\bullet}/RR_{t_2,\bullet}$. Here $v(y) = I[y = 1]$ is the indicator function and $\tau(a, b) = a/b$.

As mentioned in the previous section, in general, it is difficult to infer PATEs or RRs from observational data since the potential presence of unmeasured confounders, which affect both treatment and outcome, can bias naive estimates. With Assumption 1.1.1,

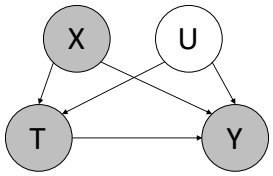


Figure 1.1: Diagrams for latent confounder models. The latent confounder model introduce a latent variable U and parameterize the effect of this latent variable on both the treatment assignment T and outcomes Y.

the following assumptions would be sufficient to identify the intervention distribution, and hence the treatment effect:

Assumption 1.2.1 (Ignorability in *do*-calculus framework) *X block all backdoor paths between T and Y [4].*

Assumption 1.2.2 (Positivity in *do*-calculus framework) *$f(T = t \mid x) > 0$ for all x.*

Assumption 1.2.1 means that X blocks every path between T and Y that contains an arrow into Y, implying that T and Y are independent given X, which is equivalent to Assumption 1.1.2 despite that they are stated under different frameworks. Likewise, Assumption 1.2.2 is the same as Assumption 1.1.3. As we mentioned previously, the strong ignorability assumption is untestable and unlikely to hold in observational studies. Therefore, sensitivity analyses, which characterize the degree to which violations of causal assumptions affect the target quantify of interest, become especially important when there exists potential unobserved confounders. We introduce U to denote the unobserved confounders that affect both the treatment and outcome (Figure 1.1).

1.3 Sensitivity Analysis

There is an extensive literature on assessing sensitivity to violations of unconfoundedness in single treatment and outcome models, dating back at least to the work of Cornfield

et al. (1959) [1] on the link between smoking and lung cancer. Since then, a wide range of strategies have been proposed for assessing sensitivity to unobserved confounding (e.g. see [12, 13, 14, 3, 15, 16, 17, 2, 18, 19]). A common strategy for sensitivity analysis is to assert that strong unconfoundedess would hold if only an additional latent variable U were observed [8]:

Assumption 1.3.1 (Latent ignorability in *do*-calculus framework) *X and U block all backdoor paths between T and Y [4].*

Assumption 1.3.2 (Latent positivity in *do*-calculus framework) *$f(T = t \mid U = u, x) > 0$ for all u and x.*

With assumptions 1.3.1 and 1.3.2, latent confounder models simultaneously specify the full conditional distributions of treatment and outcomes given both observed confounders, x, and unobserved confounders, U. The dependence of Y and T on U is indexed by a vector of sensitivity parameters $\psi = (\psi_Y, \psi_T)$. Practitioners can then reason about how assumptions about these parameters translate to different causal conclusions. Often, this is done through *calibration*, by determining reasonable ranges for ψ using analogies about observable associations and through a *robustness* assessment, by examining how strong associations with unobserved confounders must be for conclusions to change.

Concretely, in a typical latent confounder analysis, we posit densities $f(u \mid x)$, the marginal density of the latent confounders, $f_{\psi_T}(t \mid x, u)$, the conditional density (or PMF) for treatment assignment given all confounders and $f_{\psi_Y}(y \mid x, u, t)$, the outcome density in treatment arm t. The sensitivity parameters encode the relationship between both the treatment and unobserved confounders and the outcome and unobserved confounders (e.g. see [3, 20]). Latent confounder models are usually parameterized so that some specific values of the sensitivity parameters ψ indicate the "no unobserved confounding" case. For example, we can take $\psi_T = 0$ to imply that $f_{\psi_T}(t \mid x, u) = f_{\psi_T}(t \mid x)$ and

$\psi_Y = 0$ to imply that $f_{\psi_Y}(y \mid x, u, t) = f_{\psi_Y}(y \mid x, t)$. Then, when either $\psi_T = 0$ or $\psi_Y = 0$, U is not a confounder [9]. Without loss of generality, we suppress conditioning on x throughout the remainder of the manuscript, and comment on the role of observed covariates where appropriate.

A key principle of sensitivity analysis is that the observed data densities should be invariant to the sensitivity parameters [21]. Unfortunately, this principle can easily be violated in a latent confounder analysis [2]. The crux of the problem is that, in general, none of $f(u \mid x)$, $f_{\psi_T}(t \mid x, u)$ or $f_{\psi_Y}(y \mid x, u, t)$ are nonparametrically identifiable themselves, but the observed outcome density, which is a function of these densities, is identifiable. The observed outcome density, has the following form:

$$f(y \mid T = t) = \int_{\mathcal{U}} f_{\psi_Y}(y \mid t, u) f_{\psi_T}(u \mid t) du, \text{ for all } t \text{ and } \psi \tag{1.16}$$

where $f_{\psi_T}(u \mid t)$ is the conditional density of the latent confounders given treatment level t. This contrasts with the density of the intervention distribution, which is:

$$f_{\psi}(y \mid do(t)) = \int_{\mathcal{U}} f_{\psi_Y}(y \mid t, u) f(u) du \tag{1.17}$$

$$= \int_{\mathcal{U}} f_{\psi_Y}(y \mid t, u) \left[\int f_{\psi_T}(u \mid \tilde{t}) f(\tilde{t}) d\tilde{t} \right] du \tag{1.18}$$

The intervention distribution is obtained by integrating over the marginal distribution of the unobserved confounder, $f_{\psi_T}(u) = \int f_{\psi_T}(u \mid \tilde{t}) f(\tilde{t}) dt$, whereas the observed data distribution in 1.16 is obtained by integrating with respect to the conditional confounder distribution, $f_{\psi_T}(u \mid t)$. A fundamental challenge of sensitivity analysis is to specify a class of densities $f_{\psi_Y}(y \mid t, u)$ and $f_{\psi_T}(u \mid t)$, with interpretable parameters $\psi = (\psi_Y, \psi_T)$, for which the intervention distribution necessarily varies with ψ but for which the observed data distribution does not.

In fact, several authors have proposed latent variable sensitivity models with unidentified sensitivity parameters in simple settings with a single treatment and binary or categorical outcomes [8, 22, 14, 23]. A more general solution was proposed by Zhang and Tchetgen (2019) [24] who propose a semi-parametric sensitivity model in which the distribution of U is left unrestricted. Franks et al. (2019) [2] propose an alternative framework for sensitivity analysis without directly introducing latent variables by directly parameterizing the effect of the outcome on the treatment assignment. Cinelli and Hazlett (2019) [18] and Cinelli et al. (2019) [25] use moment arguments to derive confounding bias in the linear regression setting and introduce an approach for sensitivity parameter calibration based on the proportion of variance explained by the latent confounder. In the multiple treatment and/or outcome setting, additional observable implications can complicate sensitivity analysis, and thus new strategies are needed.

Unfortunately, many existing sensitivity analysis approaches perturb the observable predictions of the model and degrade the quality of observed data predictions. In the following section, we introduce the latent confounder model and discuss this difficulty in more details.

1.4 Latent Confounder Models

Latent confounder models are one of prevailing approaches for sensitivity analysis in causal inference [8]. In this section, we introduce the latent confounder model using the potential outcome framework. First of all, we restate the latent strong ignorability assumption under the potential outcome framework.

Assumption 1.4.1 (Latent ignorability in potential outcome framework)

$$[Y(0), Y(1)] \perp\!\!\!\perp T \mid X, U. \tag{1.19}$$

Assumption 1.4.2 (Latent positivity in potential outcome framework)

$$0 \leq P(T = 1 \mid X, U) \leq 1 \tag{1.20}$$

With Assumptions 1.4.1 and 1.4.2, the latent confounder model simultaneously specifies the conditional distributions of the treatment and the potential outcomes given X and U (see Figure 1.1). For example, for a binary treatment, the model can be specified as:

$$U \mid X \sim f(U \mid X) \tag{1.21}$$

$$T \mid U, X \sim f_{\psi_T}(T \mid X, U) \tag{1.22}$$

$$f_{\psi_T}(T \mid X, U) = \text{Bern}(e_{\psi_T}(X, U)) \tag{1.23}$$

$$Y(t) \mid U, X \sim f_{\psi_{Y_t}}(Y(t) \mid X, U) \tag{1.24}$$

where $f(U \mid X)$ is the density of the latent confounder; $f_{\psi_{Y_t}}(Y(t) \mid X, U)$ is the potential outcome density for treatment t given the observed and latent variables; and $e_{\psi_T}(X, U)$ is the probability of receiving treatment given both X and U, written with $e(\cdot)$ to invoke a parallel to the propensity score. The sensitivity parameters $\psi = (\psi_T, \psi_{Y_t})$ encode how the treatment and potential outcomes depend on the unobserved confounder. Despite their intuitive appeal, latent confounder models often imply observed distributions that depend on sensitivity parameters:

$$f^{\text{obs}}_{\psi,t}(Y(t) \mid X, T = t) = \int_{\mathcal{U}} f_{\psi_{Y_t}}(Y(t) \mid X, U) f_{\psi_T}(U \mid X, T = t)\, dU. \tag{1.25}$$

The above equation shows that the distribution of observed outcomes is a mixture over the mixing measure $f_{\psi_T}(U \mid X, T = t)$, and it depends on the sensitivity parameter ψ_T and ψ_{Y_t} via the mixture weights and mixture components respectively. Therefore,

when tuning the sensitivity parameters, the observed data distribution would also change correspondingly. This problem can be seen clearly in the following simple example:

Example 1.4.1 (Gaussian outcome, binary confounder and treatment) *Consider the case where outcome is continuous and there is no covariates. Assume that treatment was randomly assigned according to a Bernoulli design, but it is plausible that there exists a latent class that confounds the study. To test the robustness of our causal conclusions to the presence of such a latent class, we propose a sensitivity analysis by introducing a binary latent confounder. The model is parameterized as follows:*

$$\begin{aligned} U &\sim Bern(\xi_u), \\ T \mid U &\sim Bern(g(\alpha + \psi_T U)), \\ Y(t) \mid U &\sim N(\mu_t + \psi_{Y_t} U, \sigma^2). \end{aligned}$$

We let $h_{\psi_T} := P(U = 1 \mid T)$. According to Equation 1.25, the distribution of observed outcomes is a two-component mixture of normals for $t \in \{0, 1\}$:

$$Y(t) \mid T = t \sim h_{\psi_T} N(\mu_t + \psi_{Y_t}, \sigma^2) + (1 - h_{\psi_T}) N(\mu_t, \sigma^2), \tag{1.26}$$

where the mixture weights depend on the sensitivity parameters ψ_T, and one of the mixture components depends on the sensitivity parameter ψ_{Y_t}. The existence of such sensitivity parameters, which intervenes the observed data distribution, blurs the line between sensitivity analysis and model checking, ending up inadvertently perturbing the fit of the model and thus degrade the quality of observed data predictions.

To overcome this difficulty, we propose new methods to sensitivity analysis in the rest of this dissertation. In Chapter 2, we reparameterize the latent confounder models and decompose the effect of latent confounder, U, into confounding variations and non-

confounding variations. In Chapter 3, we utilize the Tukey's factorization and extend the Tukey's sensitivity analysis method [2] to cases where treatments are ordinal variables with more than two levels. Moreover, in order to handle modern problems, which are often of multiple concurrent treatment variables and/or high dimensional outcomes, we develop copula-based sensitivity analysis for cases with multiple treatments and outcomes in Chapter 4 and 5 respectively.

Chapter 2

Sensitivity Analysis with Reparameterized Latent Confounder Models

In this chapter, we describe a sensitivity analysis approach that leaves the observed data distribution untouched by reparameterizing latent confounder models, where the residual variance of the outcome is explicitly partitioned into the confounding and non-confounding parts. For this method, we focus on the use of potential outcome framework.

2.1 Reparameterization

The goal of our sensitivity analysis is to find sets of model specification for $Y \mid T, U, X$ and $U \mid T, X$ for which the observed data distribution, $f^{\text{obs}}_{\psi,t}(Y(t) \mid T = t, X)$, remains unchanged. To achieve this, we decompose the the effect of U in the latent confounder model (Figure 1.1) into a confounding variation, denoted by W, and non-confounding variation due to mediators and/or measurement error, denoted by E, which leads to a

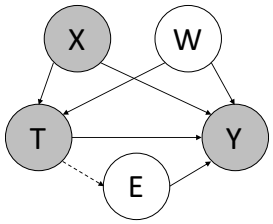


Figure 2.1: Diagram for reparameterized latent confounder model. Decomposing the effect of U into confounding variation, represented by W, and non-confounding variation due to mediators and/or measurement error, represented by E.

clean separation between the identified components and unidentified components in the specified model (Figure 2.1).

We use the following example to demonstrate our method, where both conditional confounder and outcomes are Gaussian.

Example 2.1.1 (Reparameterized Latent Confounder Model) *Suppose we are interested in a study with a continuous outcome and a binary treatment. We introduce a Gaussian variable W, which represents the shared variation due to unobserved confounders given the treatment, and a Gaussian variable E, which denotes the unshared variation due to mediators (i.e., indirect treatment effect) and/or measurement error of the outcomes. Without loss of generality, the conditional variance of W is assumed to be one, which can be easily achieved by standardization. We assume that the potential outcome $Y(t)$ is linear in μ_{X_t}, W and E, where μ_{X_t} denotes the overall mean of potential outcome $Y(t)$ for $t \in \{0, 1\}$. The model is specific as follows:*

$$Y(t) = \mu_{X_t} + \gamma W + E, \tag{2.1}$$

$$W \mid T = 0 \sim N(0, 1), \tag{2.2}$$

$$W \mid T = 1 \sim N(\phi, 1), \tag{2.3}$$

$$E \sim N(0, \sigma^2) \tag{2.4}$$

where γ and ϕ are sensitivity parameters.

In Example 2.1.1, by integrating out the W and E, we can derive the missing and observed potential outcome distributions as

$$Y(0)^{obs} \mid T = 0, X \sim N(\mu_{X_0}, \gamma^2 + \sigma^2), \tag{2.5}$$

$$Y(1)^{obs} \mid T = 1, X \sim N(\mu_{X_1} + \gamma\phi, \gamma^2 + \sigma^2), \tag{2.6}$$

$$Y(0)^{mis} \mid T = 1, X \sim N(\mu_{X_0} + \gamma\phi, \gamma^2 + \sigma^2), \tag{2.7}$$

$$Y(1)^{mis} \mid T = 0, X \sim N(\mu_{X_1}, \gamma^2 + \sigma^2). \tag{2.8}$$

Importantly, note that $E(Y(0)^{obs} \mid T = 0, X) = \mu_{X_0}$, $E(Y(1)^{obs} \mid T = 1, X) = \mu_{X_1} + \gamma\phi$, and the variance of potential outcome distributions, $\gamma^2 + \sigma^2$, are all identifiable from the observed data. Let $\tilde{\mu}_{X_{00}} := E(Y(0)^{obs} \mid T = 0, X)$ and $\tilde{\mu}_{X_{11}} := E(Y(1)^{obs} \mid T = 1, X)$. With $\tilde{\mu}_{X_{00}}$ and $\tilde{\mu}_{X_{11}}$, the missing potential outcome distributions can be alternatively written as:

$$Y(0)^{mis} \mid T = 1, X \sim N(\tilde{\mu}_{X_{00}} + \gamma\phi, \gamma^2 + \sigma^2), \tag{2.9}$$

$$Y(1)^{mis} \mid T = 0, X \sim N(\tilde{\mu}_{X_{11}} - \gamma\phi, \gamma^2 + \sigma^2). \tag{2.10}$$

Given $\gamma\phi$, the marginal contrast estimands of our interest are all identifiable as they can be written as functions of potential outcome distributions. Remarkably, we can achieve the same fixed level of confounding with any sets of sensitivity parameters (γ_1, ϕ_1) and (γ_2, ϕ_2) satisfying $\gamma_1\phi_1 = \gamma_2\phi_2$ by either having 1) very large imbalance ϕ_1 but very small regression coefficient γ_1 or 2) very small imbalance ϕ_2 but very large regression coefficient γ_2.

To generalize our idea, we can consider classes of generalized linear mixed model for the observed potential outcomes [26], which allow extra error components in the linear

predictors of generalized linear model. The distribution of these random components is not restricted to be normal and can come from an arbitrary distribution. To take advantage of the exponential family [27], we can specify the distribution of those random effects to be the conjugate of the outcome distributions. For instance, when conditional outcomes follow Poisson distribution and conditional confounders follow Gamma distribution, by integrating out the random effects, the marginal distribution of outcomes would follow Negative Binomial distribution. As another example, we can alternatively specify the distribution of conditional outcomes to be Beta and conditional confounder to be Binomial so that the marginal distribution of outcomes would follow the Beta Binomial distribution. Like we've seen in Example 2.1.1, there are two main advantages to specify models this way: first, the distribution of treatment given unobserved confounder will be logistic in the sufficient statistic for the conditional distribution of the confounder, which makes the calibration of sensitivity parameters more intuitive and interpretable; second, the observed outcome distribution $f^{\text{obs}}_{\psi,t}(Y(t) \mid T = t, X)$ will be compound distributions that have been well studied in the literature.

2.2 Calibration

2.2.1 Calibrating the Sensitivity Parameter γ

We can calibrate the magnitude of γ by considering the partial R^2 of W in the observed potential outcome model, specifically

$$R^2_{Y(t)\sim W|X,T} = \frac{\text{SSR}_{Y(t)\sim X,T} - \text{SSR}_{Y(t)\sim W,X,T}}{\text{SSR}_{Y(t)\sim X,T}} = \frac{\gamma^2}{\gamma^2 + \sigma^2}, \tag{2.11}$$

where $\text{SSR}_{Y(t)\sim X,T}$ and $\text{SSR}_{Y(t)\sim W,X,T}$ denote the residual sum of square of the models by regressing $Y(t)$ on X, T and W, X, T respectively. $R^2_{Y(t)\sim W|X,T}$ represents the fraction

of variation unexplained by X and T in $Y(t)$ that can be explained by adding W into the previous model.

Based on expert knowledge, we may set $R^2_{Y(t)\sim W|X,T}$ to some reasonable $R^2_{Y(t)\sim X_j|X_{-j},T}$, which stands for the fraction of additional variation in $Y(t)$ explained by adding X_j into the model with all other covariates X_{-j} and treatment T. This is motivated by the idea that the information gained by adding W to X and T as predictors of the potential outcome model is comparable to the information gained by adding X_j to X_{-j} and T.

2.2.2 Calibrating the Sensitivity Parameter ϕ

From Equation 2.2 and 2.3, we can deduce that

$$\frac{P(T=1 \mid W)}{P(T=0 \mid W)} \propto \frac{\pi_T}{1-\pi_T} \exp\left(-\phi W\right) \tag{2.12}$$

where $\pi_T := P(T=1)$. From Equation 2.12, we can see that the conditional treatment variable is logistic in W. We therefore posit the treatment assignment model:

$$T \mid X, W \sim \text{Bern}(\text{logit}^{-1}(\alpha(X) - \phi W)), \tag{2.13}$$

where the log-odds of receiving treatment are linear in W, with $\alpha(X)$ being a function of X and $\text{logit}^{-1}(x) = (1+\exp(-x))^{-1}$. One of the advantages gained from the treatment assignment specification 2.13 is that the sensitivity parameter ϕ has a very natural and intuitive interpretation, which describes how treatment assignment depends marginally on the confounding part, W.

To calibrate the sensitivity parameter ϕ, we adopt the idea of "implicit R^2" from Imbens (2003) [3], denoted by ρ^2 in the following. Similar to the calibration of γ, we consider the implicit R^2, $\rho^2_{T\sim W|X}$, which stands for the fraction of residual variance of T

after conditioning on X that can be explained by W. We may set $\rho^2_{T\sim W|X}$ to $\rho^2_{T\sim X_j|X_{-j}}$ with belief that the variation in T explained by adding W into the model with X already included is comparable to the variation explained by adding X_j to the model with X_{-j} already included.

2.3 Application: Analysis of NHANES data

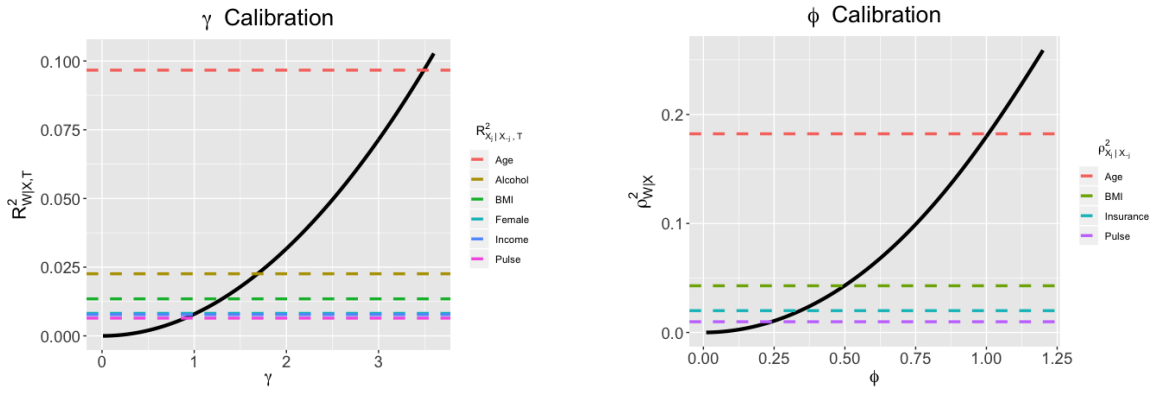
We illustrate our reparameterized method using the data from the Third National Health and Nutrition Examination Survey (NHANES III) (Center for Disease Control and Prevention (CDC), 1997), we aim to estimate the effect of "taking two or more anti-hypertensives" on average diastolic blood pressure. We follow the settings of Dorie et al. (2016) [20] and Frankset al. (2018) [28], and utilize pre-treatment covariates like race, gender, age, income, body mass index (BMI), and etc. Therefore, in our case, $Y(t)$ corresponds to the average diastolic blood pressure for a subject in treatment arm t, where $t = 1$ indicates the subject was taking two or more anti-hypertensive medications and $t = 0$ indicates the subject was not.

We assume that the underlying data generating process follows Example 2.1.1. Following Dorie et al. (2016) [20], we first fit the observed response surface, $f_t^{obs}(Y(t) \mid T = t, X) \sim N(\tilde{\mu}_{X_{tt}}, \sigma_t^2)$, using a flexible nonparametric method, called Bayesian Additive Regression Tree (BART), with R package BART [29]. By Esquation 2.5-2.8, the treatment effects can be expressed in terms of $\mu_{X_{00}}$ and $\mu_{X_{11}}$ as

$$\text{PATT:} \quad \tilde{\mu}_{X_{11}} - (\tilde{\mu}_{X_{00}} + \gamma\phi), \tag{2.14}$$

$$\text{PATC:} \quad (\tilde{\mu}_{X_{11}} - \gamma\phi) - \tilde{\mu}_{X_{00}}, \tag{2.15}$$

$$\text{PATE:} \quad \frac{N_1}{N} * \text{PATT} + \frac{N_0}{N} * \text{PATC}, \tag{2.16}$$



(a) γ vs $R^2_{Y(t)\sim W|X,T}$ (b) ϕ vs $\rho^2_{T\sim W|X}$

Figure 2.2: Calibration of the analysis for the NHANES data. (a) The magnitude of the sensitivity parameter γ increases with the residual coefficient of determination, $R^2_{Y(t)\sim W|X,T}$. For comparison, we also mark the partial coefficients of variation for some important predictors by dashed lines and calibrate the magnitude of γ based on the most important covariate, the "Age". (b) The magnitude of the sensitivity parameter ϕ increases with the implicit residual coefficient of determination, $\rho^2_{T\sim W|X}$. For comparison, we mark the implicit partial coefficients of variation for some important predictors by dahsed lines and also calibrate the magnitude of ϕ based on covariate "Age".

where N is the total number of observations in the dataset, N_0 and N_1 denote the number of observations in the control and treatment groups respectively. Thus, treatment effects 2.14-2.16 can be identified if the product of sensitivity parameters $\gamma\phi$ is known.

Next, we calibrate the magnitude of the sensitivity parameters using the approach outlined in Section 2.2. According to figure 2.2a, "Age" has largest partial R^2 in the outcome model, which is around 0.095, we limit the magnitude of sensitivity parameter γ accordingly so that $\mid \gamma \mid \leq 3.5$. Similarly, we calibrate ϕ using the implicit R^2 in the treatment assignment model, and limit $|\phi|$ up to 1 according to the largest $\rho^2_{T\sim X_j|X_{-j}}$, which is $\rho^2_{T\sim Age|-Age} \approx 0.18$.

We visualize our ATE estimates for a grid of sensitivity parameters in Figure 2.3. "NS" denotes "not significant", by which we mean that 95% posterior credible interval of the ATE contains 0. Under unconfoundedness ($\gamma = \phi = 0$), the posterior for the ATE

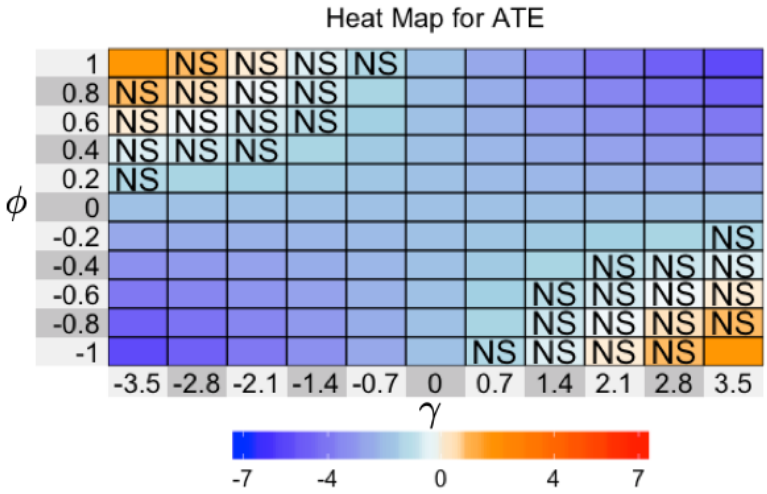


Figure 2.3: Average treatment effect for different settings of γ and ϕ in the analysis of NHANES data. NS denotes "not significant". Under unconfoundedness, the ATE is negative and significant. The conclusion can vary a lot by the setting of sensitivity parameters.

is approximately -1.88 mmHg, there is enough posterior uncertainty that the effect is significant different from 0 (light blue box). When sensitivity parameters γ and ϕ have the same sign, located along the bottom left to upper right in figure 2.3, the treatment effect will be amplified (dark blue box), which means the average diastolic blood pressure can be decreased to a larger degree by "taking two or more anti-hypertensives". In contrast, when sensitivity parameters γ and ϕ have opposite signs, located along the upper left to bottom right in figure 2.3, the treatment effect will be reduced and even reversed in extreme cases (orange box) where "taking two or more anti-hypertensive" will be harmful and surprisingly increase the average diastolic blood pressure.

Chapter 3

Sensitivity Analysis via Tukey's Factorization with Ordinal Treatments

In this chapter, we introduce our second method that can separate sensitivity analysis from model checking following the framework proposed by Franks et al. (2019) [28]. This approach directly models the dependence of the treatment assignment mechanism on the potential outcomes instead of explicitly introducing a latent variable and parameterizing the effect of this latent variable on both the treatment assignment and outcomes.

3.1 Tukey's Factorization

The foundation of the approach of Franks et al. (2019) [28] is based on a unique factorization introduced in the missing data literature, known as Tukey's factorization [30] or the extrapolation approach [31]. This approach can be applied to observational causal inference by writing the complete data in terms of the observed data densities,

the treatment assignment mechanism, and a term capturing the conditional dependence between potential outcomes. From Franks et al. (2019) [28], the complete data density can be written as

$$f(T, [Y(0), Y(1)] \mid X) = \prod_{t=0}^{1} \left[f^{obs}(Y(t) \mid T = t, X) f(T = t.X) \frac{f^{asgn}(T \mid Y(t), X)}{f^{asgn}(T = t \mid Y(t), X} \right] \cdot \frac{1}{f(T \mid X)} \cdot c(F(Y(0) \mid T), F(Y(1) \mid T) \mid T, X), \tag{3.1}$$

where the conditional copula, $c(F(Y(0) \mid T), F(Y(1) \mid T) \mid T, X)$, is defined as $\frac{f(Y(0),Y(1)|T,X)}{f(Y(0)|T,X)f(Y(1)|T,X)}$.

A major advantage of this approach is that the observed outcome distribution $f^{obs}(Y(t) \mid T = t, X)$ can be identified using non-parametric or flexible machine learning method and the treatment assignment $f^{asgn}(T \mid Y(t), X)$, which is not identifiable but interpretable. Additionally, as Franks et al. (2019) [28] point out, although $c(F(Y(0) \mid T), F(Y(1) \mid T) \mid T, X)$ is not identifiable from the observed data, it is not necessary for estimating many common causal estimands. It is clear to see that, in Equation 3.1, the observed outcome distribution is distinctly separated from the unidentifiable terms, $f^{asgn}(T \mid Y(t), X)$ and $c(F(Y(0) \mid T), F(Y(1) \mid T) \mid T, X)$. Therefore, there is a clean separation between sensitivity analysis and model checking.

In Franks et al. (2019), they only consider the scenario where all variables are univariate, however, the causal inference problems with multivariate treatments are quite common in practice. For example, a medical researcher may wish to find out what the treatment effects of a new drug with low, median and high dosage are respectively. Here, instead of binary treatment, we consider categorical treatment with ordinal levels. However, the factorization 3.1 only applies to the standard setting where T is binary and $Y(t) \in \mathcal{R}$ for $t \in \{0, 1\}$. This motivates us to extend the extrapolation framework to

cases with ordinal treatment of multiple levels. In the next section, we show our extension of the Tukey's factorization of the complete data density in the ordinal treatment case.

3.2 Generalizing Tukey's Factorization

Following the notations introduced in Section 1.1, let T denote a categorical treatment variable with ordinal levels $1, \cdots, M$ $(M > 2)$, and $Y(t)$ denote the potential outcomes of the treatment level t, $t \in \{1, \cdots, M\}$.

First and foremost, we clarify the definition of the estimands of our interest in the multivariate treatment cases with potential outcome notations. Following the detailed discussion of the causal estimands with multiple treatments by Lechner (2001) [32], we focus on pair-wise comparisons of the treatment effect between the treatment t and s:

$$\gamma^{t,s} = E(Y(t) - Y(s)) = E(Y(t)) - E(Y(s)), \tag{3.2}$$

$$\alpha^{t,s} = E(Y(t) - Y(s) \mid T = s, t) = E(Y(t) \mid T = s, t) - E(Y(s) \mid T = s, t), \tag{3.3}$$

$$\theta_t^{t,s} = E(Y(t) - Y(s) \mid T = t) = E(Y(t) \mid T = t) - E(Y(s) \mid T = t). \tag{3.4}$$

$\gamma^{t,s}$ denotes the expected effect of treatment t relative to treatment s for a participant drawn randomly from the population. Similarly, $\alpha^{t,s}$ and $\theta_t^{t,s}$ respectively denotes the same effect for a participant randomly selected from the group of participants participating in either s or t, and t only. Note that $\gamma^{t,s}$ and $\alpha^{t,s}$ are weighted combination of estimands $\theta^{t,s}$, we therefore focus on the estimation of $\theta^{t,s}$ in the following. To estimate $\theta_t^{t,s}$, we need to know both the observed potential outcomes distribution $f(Y(t) \mid T = t)$ and the unobserved potential outcomes distribution $f(Y(s) \mid T = t)$. The estimation of the observed outcome density is trivial, while it is challenging to estimate the missing

outcome distributions, which is the focus of our following discussion.

We extend Tukey's factorization of the complete data density 3.1 to the multi-level treatment case as

$$f(T, [Y(1), \cdots, Y(M)] \mid X) = \prod_{t=1}^{M} \left[f^{obs}(Y(t) \mid T = t, X) f(T = t \mid X) \frac{f^{asgn}(T \mid Y(t), X)}{f^{asgn}(T = t \mid Y(t), X} \right] \cdot \left(\frac{1}{f(T \mid X)} \right)^{M-1} \cdot c(F(Y(1) \mid T), \cdots, F(Y(M) \mid T) \mid T, X), \tag{3.5}$$

where the conditional copula is defined similarly to the binary case,

$$c(F(Y(1) \mid T), \cdots, F(Y(M) \mid T) \mid T, X) := \frac{f(Y(1), \cdots, Y(M) \mid T, X)}{f(Y(1) \mid T, X) \cdots f(Y(M) \mid T, X)}. \tag{3.6}$$

In above equation, the joint density is decomposed into the observed data densities $f^{obs}(Y(t) \mid T = t, X)$, the treatment assignment mechanisms $f^{asgn}(T \mid Y(t), X)$ and the conditional copula, $c(F(Y(1) \mid T), \cdots, F(Y(M) \mid T) \mid T, X)$, capturing the conditional dependence between potential outcomes. Same as the binary case, the Tukey's factorization of the complete data in the multi-level treatment case also leaves the observed outcome distribution $f^{obs}(Y(s) \mid T = s, X)$ free of the sensitivity parameter γ_s, which leads to a clean separation between model checking and sensitivity analysis.

Notably, Equation 3.5 implies that the missing outcome distribution $f^{mis}(Y(s) \mid T = t, X)$ is a tilt of the observed outcome distribution $f^{obs}(Y(s) \mid T = s, X)$,

$$f^{mis}(Y(s) \mid T = t, X) \propto f^{obs}(Y(s) \mid T = s, X) \frac{f^{asgn}_{\gamma_s}(T = t \mid Y(s), X)}{f^{asgn}_{\gamma_s}(T = s \mid Y(s), X)}. \tag{3.7}$$

Equation 3.7 shows that the missing outcome distribution depends on the observed outcome distribution and the treatment assignment. While the observed outcomes distribution can be identified using non-parametric or flexible machine learning methods such

as Bayesian Regression Tree Model (BART) [33], the treatment assignment mechanism, parameterized by the sensitivity parameter γ_s, is not identifiable but can be easily reasoned out (discussed in Section 3.3). Hence, given the sensitivity parameter γ_s, we would be able to identify the marginal contrast estimand $\theta_t^{t,s}$, which is a function of observed and missing outcome distributions.

3.3 Logistic Selection with Exponential Family Models

In this section, we discuss practical implementation of the Tukey's method in the two most interested cases where outcome variables are binary or Gaussian. For the specification of the treatment assignment model, in order to facilitate the calibration and interpretation of the sensitivity parameters, we assume that the adjacent-categories logits are linear in some sufficient statistics of the potential outcomes, and, thus, the sensitivity parameters describe how treatment assignment depends marginally on each potential outcome. One of the difficulties for sensitivity analyses in multivariate treatment settings is that the number of sensitivity parameters increases substantially with the dimension of the treatment variable. We address this issue by considering proportional odds with ordinal treatments, which largely cut down the number of sensitivity parameters to one in each treatment arm.

Let's first consider the case where outcomes are binary.

Example 3.3.1 (Binary Outcomes) *Assuming that the observed potential outcomes in the treatment arm s follow the Bernoulli distribution,*

$$f^{obs}(Y(s) \mid T = s) \sim Bern(logit^{-1}(\mu_s(x))), \qquad s = 1, \cdots, M. \tag{3.8}$$

and specifying the treatment assignment model using adjacent-categories logits model in which the adjacent-categories logits are linear in the sufficient statistics of Bernoulli distribution for the outcomes in the treatment arm s, i.e., $Y(s)$, as

$$\log \frac{p(T = j \mid Y(s), X)}{p(T = j + 1 \mid Y(s), X)} = \alpha_{js} + \beta_s(x) + \gamma_s Y(s), \qquad j = 1, \cdots, M - 1. \tag{3.9}$$

In the above Example 3.3.1, we can derive the missing potential outcome distribution of potential outcome $Y(s)$ in the treatment arm t based on Equation 3.7, which equals

$$f^{mis}(Y(s) \mid T = t, X) \sim \text{Bern}(\text{logit}^{-1}(\mu_s(x) + (s - t)\gamma_s)), \qquad t \neq s,\ t = 1, \cdots, M. \tag{3.10}$$

Equation 3.10 shows that the log-odds of the missing potential outcomes is an additive shift to the observed potential outcomes. Therefore, we are able to calculate any marginal contrast causal estimands given the sensitivity parameter γ_s.

Beside the binary outcome, our Tukey's method also applies well to cases with Gaussian outcomes.

Example 3.3.2 (Gaussian Outcomes) *Assume that the observed potential outcomes in the treatment arm s follow the normal distribution:*

$$f^{obs}(Y(s) \mid T = s, X) \sim N(\mu_s(x), \sigma_s^2), \qquad s = 1, \cdots, M. \tag{3.11}$$

In the treatment assignment model, let adjacent-categories logits be linear in the sufficient statistics of normal distribution for the outcomes in the treatment arm s, i.e., $Y(s)$ and

$Y^2(s)$:

$$\log \frac{p(T=j \mid X, Y(s))}{p(T=j+1 \mid X, Y(s))} = \alpha_{js} + \beta_s(x) + \gamma_s Y(s) + \psi_s Y^2(s), \qquad j = 1, \cdots, M-1. \tag{3.12}$$

Under the specification of Example 3.3.2, the missing potential outcome distribution for outcome $Y(s)$ in the treatment t can be derived according to Equation 3.7 as

$$f^{mis}(Y(s) \mid T=t, X) \sim N(\frac{\mu_s(x) + (s-t)\gamma_s\sigma_s^2}{1-2(s-t)\psi_s\sigma_s^2}, \frac{\sigma_s^2}{1-2(s-t)\psi_s\sigma_s^2}),\ t \neq s,\ t = 1, \cdots, M, \tag{3.13}$$

which implies that the distribution of missing potential outcomes is within the same exponential family as of the observed outcomes. Similar to the binary outcome case, given the sensitivity parameters γ_s and ψ_s, we are able to identify the causal estimands of our interest.

3.4 Calibration

As of now, we decently specify a set of causal models indexed by sensitivity parameters that are decoupled from the observed data model. Given values of sensitivity parameters, we are able to deduce the causal effects accordingly. The next natural question we would ask is how to find plausible regions for these sensitivity parameters exploiting the expert knowledge? So, in this section, we propose a strategy to solve this problem. Our key idea is that calibrating the magnitude of sensitivity parameters to the amount of variation in the treatment assignment T that is explained by $Y(t)$ with respect to what is counted for X.

Considering the generalized propensity score model, which models the probability of treatment assignment T over the observed covariates X by adjacent-categories logits

model,

$$\log \frac{p(T = j \mid X)}{p(T = j + 1 \mid X)} = c_{js} + m_s(x), \qquad j = 1, \cdots, M - 1. \tag{3.14}$$

Note that this generalized propensity score model is identifiable by the observed data. Under model 3.14, we can measure the proportion of variation in T explained by X using the implicit R^2 proposed by McKelvey and Zavoina (1975) [34] as

$$\rho^2_X := \frac{Var(m_s(x))}{Var(m_s(x)) + \pi^2/3}. \tag{3.15}$$

Accordingly, we define the fraction of previously unexplained variation in T that can be explained by adding one of the observed predictors X_j to other predictors X_{-j} in parallel with ordinary partial R^2 as

$$\rho^2_{X_j|X_{-j}} = \frac{\rho^2_X - \rho^2_{X_{-j}}}{1 - \rho^2_{X_{-j}}}. \tag{3.16}$$

To calibrate the sensitivity parameters in the binary outcome case (Example 3.3.1), we comparatively consider the variation in T explained by the treatment assignment model 3.9, which can be measured as

$$\rho^2_{X,Y(s)} = \frac{Var(\beta_s(x) + \gamma_s Y(s))}{Var(\beta_s(x) + \gamma_s Y(s)) + \pi^2/3} \tag{3.17}$$

Similar to $\rho^2_{X_j|X_{-j}}$, the fraction of previously unexplained variation in T that can be explained by adding the potential outcome $Y(t)$ to the observed predictors X can be measured as

$$\rho^2_{Y(s)|X} = \frac{\rho^2_{Y(s),X} - \rho^2_X}{1 - \rho^2_X}, \tag{3.18}$$

which is a function of the sensitivity parameter γ_s. Therefore, we can calibrate $\rho^2_{Y(s)|X}$ with respect to the identifiable quantity $\rho^2_{X_j|X_{-j}}$ in order to find the plausible region for the sensitivity parameter γ_s.

3.5 Simulation

In this section, we demonstrate our Tukey's sensitivity analysis approach by simulation following the design in Gu et al. (2019) [35].

We consider categorical treatment T with three ordinal levels 1, 2, 3, binary potential outcomes, and ten predictors, three unobserved confounders U_1, U_2, U_3 and seven observed covariates $X_1, \cdots, X_7$, which are independent and identically distributed by the standard normal distribution $N(0,1)$. Our goal is to estimate the treatment effect $\theta_2^{2,1} = E(Y(2) - Y(1) \mid T = 2)$.

The treatment T is generated from the model,

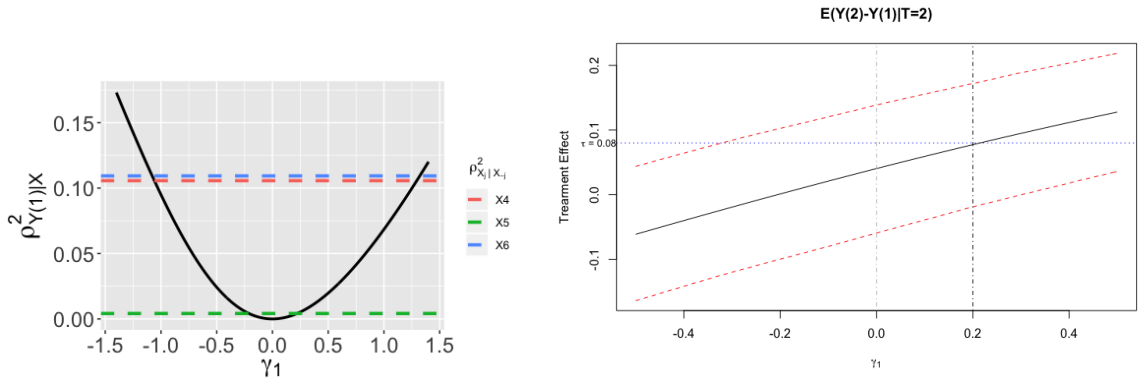
$$\log \frac{P(T=1 \mid X,U)}{P(T=2 \mid X,U)} = \alpha_1 + \eta(X,U),$$
$$\log \frac{P(T=2 \mid X,U)}{P(T=3 \mid X,U)} = \alpha_2 + \eta(X,U),$$

where $\eta(X,U) = 0.5U_1 + 0.7U_2 + 0.5U_3 + 0.8X_4 + 0.2X_5 + 0.8X_6$. We let $\alpha_1 = 0.6$ and $\alpha_2 = 1.7$ so that the ratio of units in three treatment groups, $N_1 : N_2 : N_3$, equals $4 : 2 : 1$. In addition, the response surface is assumed as

$$E(Y(1) \mid X,U) = P(Y(1)=1 \mid X,U) = \text{logit}^{-1}\{\xi(X,U)\}/2.5,$$
$$E(Y(2) \mid X,U) = P(Y(2)=1 \mid X,U) = \text{logit}^{-1}\{\xi(X,U)\}/2.5 + \tau,$$
$$E(Y(3) \mid X,U) = P(Y(3)=1 \mid X,U) = \text{logit}^{-1}\{\xi(X,U)\}/2.5 + 2\tau,$$

where $\tau = 0.08$ and $\xi(X,U) = 0.8U_1 + 0.6U_2 + 0.8U_3^2 + 0.5X_4 + 0.7X_5 + 0.9X_6^2 + 0.6U_1X_4$. Based on the model specification of our simulation, the true value of the estimand of our interest, $\theta_2^{1,2}$, is equal to τ.

We calculate $\theta_2^{1,2}$ and calibrate the sensitivity parameter γ_1 based on our previous



(a) $\rho^2_{Y(1)|X}$ vs γ_1 (b) estimated $\theta_2^{2,1}$

Figure 3.1: (a) $\rho^2_{Y(1)|X}$ vs γ_1. $|\ \gamma_1\ |$ increases with the implicit partial variance explained by $Y(1)$, $\rho^2_{Y(1)|X}$. For comparison, we mark the partial variance explained by observed covariates with non-zero effect by horizontal dashed lines, and calibrate the magnitude of γ_1 based on X_6. (b) Estimated $\theta_2^{2,1}$, the expected effect of $T = 2$ relative to $T = 1$ for units drawn randomly from the the treatment arm 2 with different assumptions about the sensitivity parameter γ_1. Under the unconfoundedness ($\gamma_1 = 0$), the estimated causal effect is around 0.04. After accounting for an amount of confoundedness corresponding to $\gamma_1 = 0.2$, the estimated causal effect would be closer to the truth of $\theta_2^{1,2}$, 0.08.

discussion of this chapter. In Figure 3.1a, we show how $\rho^2_{Y(1)|X}$ changes with γ_1, where $\rho^2_{Y(1)|X}$ denotes the proportion of additional variation explained by adding $Y(1)$ to the treatment assignment model with covariates X already included as predictors. To compare $\rho^2_{Y(1)|X}$ with observed quantities $\rho^2_{X_j|X_{-j}}$, we also include the partial coefficient of variation $\rho^2_{X_j|X_{-j}}$ for non-zero effect covariates, X_4, X_5 and X_6 in the plot by horizontal dashed lines.

Following the discussion in Section 3.4, we bound the magnitude of sensitivity parameters γ_1 with respect to the partial variance explained by covariate X_6, with $\rho^2_{X_6|X_{-6}} \approx 0.004$. More specifically, to find the corresponding value of γ_1, we set $\rho^2_{Y(1)|X} = 0.04$ and solve Equation 3.18 for γ_1, which turns out to be about 0.2.

We display the sensitivity analysis results in Figure 3.1b, where it shows that how $\theta_2^{2,1}$ changes to unmeasured confounders within the range of $[-0.5,\ 0.5]$ for γ_1. The black

solid line denotes the estimated average treatment effect of $\theta_2^{2,1}$, and the red dashed lines represent the upper and lower bounds of the 95% posterior credible interval at different settings of the sensitivity parameter. In this plot, we can see that the estimated treatment effect would increase from 0.04 ($\gamma_1 = 0$) to 0.1, getting closer to its true value 0.08, after adjusting for a level of confoundedness corresponding to $\gamma_1 = 0.2$.

3.6 Discussion

The two methods proposed in Chapter 2 and 3 both clearly separate the identified and the unidentified portion of the data-generating process, which allows us to evaluate the sensitivity of our causal conclusions under a range of hypothetical assumptions without re-fitting potential outcome models, and largely reduces the computational cost. The two methods have their own advantages. For the Tukey's method, it can be easily applied to a class of models, the logistic selection with mixtures of exponential families (logistic-mEF models), in contrast, there is no universal way to reparameterize the latent confounder models. On the other hand, the latent class model implies a bounded ignorance region for the treatment effect, while it is not the case with the Tukey's method.

This motivates us to develop sensitivity analysis methods that have both advantages of the two methods above, i.e., being versatile to different data types and can imply bounded ignorance regions under appropriate assumptions. Moreover, although we have successfully extend Tukey's extrapolation method to cases with ordinal treatments, modern problems often involve diverse data types of high-dimensional treatments or even outcomes. Sensitivity analysis methods for causal inference problems with high-dimensional treatments and/or outcomes are underdeveloped in the literature. Thus, in the final chapters of this dissertation, we provide sensitivity analysis methods that are applicable to high-dimensional settings of different types of treatments and outcomes by

making use of a copula factorization.

Chapter 4

Copula-based Sensitivity Analysis with Multiple Treatments

Recent work has focused on the potential and pitfalls of causal identification in observational studies with multiple simultaneous treatments. On the one hand, a latent variable model fit to the observed treatments can identify essential aspects of the distribution of unobserved confounders. On the other hand, it has been shown that even when the latent confounder distribution is known exactly, causal effects are still not point identifiable. Thus, the practical benefits of latent variable modeling in multi-treatment settings remain unclear. We clarify these issues with a sensitivity analysis method that can be used to characterize the range of causal effects that are compatible with the observed data. Our method is based on a copula factorization of the joint distribution of outcomes, treatments, and confounders, and can be layered on top of arbitrary observed data models. We propose a practical implementation of this approach making use of the Gaussian copula, and establish conditions under which causal effects can be bounded. We also describe approaches for reasoning about effects, including calibrating sensitivity parameters, quantifying robustness of effect estimates, and selecting models which are

most consistent with prior hypotheses.

4.1 Introduction

Although it is well-established that, in the conventional causal inference setting where both the treatment and outcome are single, treatment effects are not generally identifiable in the presence of unobserved confounding, recent work has focused on whether this challenge can be mitigated when there are multiple simultaneous treatments. Intuitively, dependence among multivariate treatments could provide information about latent confounders, which could in turn be leveraged to facilitate causal inference and identification. This intuition has motivated latent variable approaches such as "the deconfounder", a much discussed approach for estimating causal effects for multiple treatments [36].

Unfortunately, it was shown that this strategy has limited practical applicability for point identification and estimation of causal effects. For example, D'Amour (2019a) [37] and D'Amour (2019b)[38] note the lack of general nonparametric identification in the deconfounder approach, and show that the special cases in which the approach does provide identification correspond to situations where all confounding is already observed. Ogburn et al. (2019) [39] and Ogburn et al. (2020) [40] provide several additional counterexamples and detailed rebuttals to previous theoretical results. Even in the special cases where causal effects are identifiable, Grim- mer et al. (2020) [41] demonstrate through a suite of simulations and real-data analyses that the deconfounder cannot consistently outperform naive regression. They conclude by further arguing that the deconfounder assumptions are too strong to be applicable in practice.

These challenges are particularly relevant because similar strategies are used in genomics [42], computational neuroscience, social science and medicine [43], and time series applications [44]. Given the practical importance of causal inference with multiple treat-

ments, recent work has focused on stronger identifying assumptions for causal effects in the multi-treatment setting. Miao et al. (2020) [45] propose identifying assumptions involving instrumental variables and in settings when over half of the treatments are assumed to have a null effect while Kong et al. (2019) [46] consider identification in a parametric model with binary outcomes.

This literature has revolved around a binary question about point identification: can be causal effects be identified or not? Negative answers to this question often run counter to practitioners' intuitions in specific data analyses. In particular, it is intuitive that a latent variable model should provide *some* helpful information, even if this information is not enough to fully identify causal effects.

To address this issue, we propose that sensitivity analysis—which explores the range of causal effects that are consistent with the observed data in the context of a given problem—can resolve this tension. Specifically, sensitivity analysis can show how much is gained by leveraging latent structure in a given application, even if this (usually) falls short of fully identifying the causal effect of interest. To this end, we propose a sensitivity analysis approach to help practitioners better understand confounding in the multi-treatment setting, focusing on the special case where the conditional distribution of unobserved confounders given treatments is identifiable. To extend sensitivity analysis to the multi-treatment setting, we propose a general copula-based decomposition of standard latent variable–based sensitivity analysis models. This factorization allows us to precisely separate the parts of the model that are, and are not, identified in the multi-treatment setting.

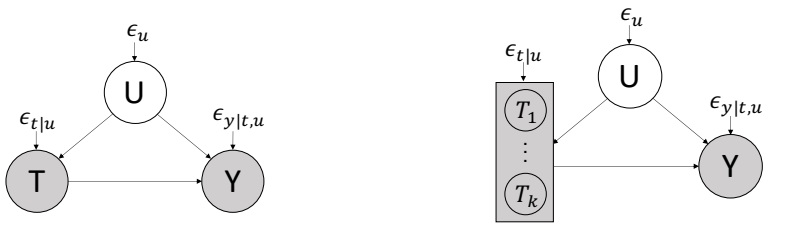
For practical analyses, we propose a specialization of the general decomposition, which specifies a sensitivity model based on invariant Gaussian copulas. While this Gaussian copula specification only covers a sub-family of sensitivity models expressible in our general formulation, we show that it captures several essential qualitative aspects of

confounding in the multi-treatment setting. In this context, we establish that there are important advantages to multi-treatment inference over single-treatment inference for characterizing sensitivity to unobserved confounding. Specifically, under appropriate assumptions, we establish that the number of effective sensitivity parameters is halved in multi-treatment inference and that this implies that the magnitude of causal effects can be bounded.

The chapter proceeds as follows. In Section 4.2 we describe our basic framework for latent variable sensitivity analysis via a copula factorization. In Section 4.3 we introduce a special case of the more general approach in which we assume confounder-outcome relationships can be characterized by a Gaussian copula. In Section 4.4 we provide some theoretical insights into bias and confounding with the Gaussian copula. We discuss sensitivity parameter interpretation, calibration, and measures of robustness in Section 4.5 and, finally, in Section 4.6 and Section 4.7 we demonstrate our approach in simulation and with the movie example analyzed by Wang and Blei (2018) [36] and later reanalyzed by Grimmer et al. (2020) [41].

4.2 Sensitivity Analysis via Copula Parameterizations with Multi-Treatment

The multiple treatment setting presents unique challenges for sensitivity analysis. In particular, the additional structure imposed in studies with multiple treatments introduces new observable implications, muddling issues of identifiability. We focus on the specific case in which the conditional confounder distribution $f_{\psi_T}(u \mid t)$ may be partially identifiable from the multiple treatments. To adapt sensitivity analysis to this setting, sensitivity parameters must not only be decoupled from the observed data distribution;



(a) Univariate Case (b) Multiple Treatments

Figure 4.1: Treatment T, outcome Y, unobserved variables U. ϵ_u, $\epsilon_{t|u}$ and $\epsilon_{y|t,u}$ are respectively the random noises of U, T and Y.

parameters describing the treatment-confounder relationships must also be decoupled from parameters describing the outcome-confounder relationships. We tackle this challenge by factorizing the joint distribution $f(y, u, t)$ using a copula, which decompose joint distributions of variables into their marginal distributions and joint dependence.

In this section, we introduce our copula-based factorization, discuss how it applies to the multiple treatment setting, and then discuss the Gaussian copula parameterization, a special case that we will use in theoretical analysis and methods development in the remainder of this chapter.

4.2.1 General Copula-Based Formulation

Our approach is based on the following factorization. The model for Y conditional on treatments and unobserved confounders can be decomposed into the observed data density and a conditional copula as

$$f_\psi(y \mid u, t) = f(y \mid t) c_{\psi_Y}(F_{Y|t}(y), F_{U|t}^{\psi_T}(u) \mid t) \tag{4.1}$$

where $F_{Y|t}$ is the CDF of $f(y \mid t)$ and $F_{U|t}$ is the CDF of $f_{\psi_T}(u \mid t)$. c_{ψ_Y} is the conditional copula density, defined on the unit hypercube and parameterized by ψ_Y, which characterizes the joint density of Y and U conditional of $T = t$ after transforming the

marginals to uniform random variables [47]. By explicitly factoring the observed outcome density, $f(y \mid t)$, out of the complete data distribution, we ensure that the left hand side of Equation 1.16 is invariant to ψ, establishing that there are no observable implications of varying the copula parameters. Moreover, this factorization holds for all densities (or PMFs) $f(y \mid t)$ and $f_{\psi_T}(u \mid t)$ and any number of treatments, and thus can be used to characterize the outcome-confounder dependence for any model of the observables.

With Equation 4.1, we can express the intervention distribution, $f_\psi(y \mid do(t))$, as:

$$f_\psi(y \mid do(t)) = f(y \mid t) \int c_{\psi_Y}(F_{Y|t}(y), F_{U|t}^{\psi_T}(u) \mid t) f(u) du, \tag{4.2}$$

$$= f(y \mid t) \int c_{\psi_Y}(F_{Y|t}(y), F_{U|t}^{\psi_T}(u) \mid t) \left[\int f_{\psi_T}(u \mid \tilde{t}) f(\tilde{t}) d\tilde{t} \right] du \tag{4.3}$$

The observed outcome distribution $f(y \mid t)$ and the marginal distribution of treatment $f(t)$ are clearly invariant to the sensitivity parameters by construction. The intervention distribution $f(y \mid do(t))$ is parameterized by ψ_Y, the parameter governing the conditional dependence between Y and U given T and by ψ_T, the parameters governing the conditional distribution of U given T.

Given $f(y \mid t)$, $f_{\psi_T}(u \mid t)$ and c_{ψ_Y} we can compute the expected value of any function of the outcome under the intervention distribution, $E[v(Y) \mid do(t)] = \int v(y) f(y \mid do(t)) dy$. This can be in turn used to compute any marginal contrast estimand. By applying Equation 4.2, we write this intervention expectation as

$$E[v(Y) \mid do(t)] \; = \int v(y) w_\psi(y, t) f(y \mid t) dy, \tag{4.4}$$

where $w_\psi(y, t) = \int c_{\psi_Y}(F_{Y|t}(y), F_{U|t}^{\psi_T}(u) \mid t) f_{\psi_T}(u) du$ is the importance weight associated with sampling from the observed data distribution instead of the intervention distribution. In practice, we can approximate the marginal distribution of the unobserved

confounder with the mixture density $f_{\psi_T}(u) \approx \frac{1}{n}\sum_i f^{\psi_T}(u \mid t_i)$ where $t_i \in \mathcal{T}$ is the ith observed treatment and $\mathcal{T}$ is the set of all observed treatment vectors. Thus, the importance weight can be approximated as

$$w_\psi(y,t) \approx \frac{1}{|\mathcal{T}|}\sum_{t_i \in \mathcal{T}}\left[\int c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u) \mid t) f_{\psi_T}(u \mid t_i) du\right]. \tag{4.5}$$

We use this approximation to derive importance sampling algorithm for computing the the expected value in Equation 4.4 for any copula and conditional confounder distributions $f(u \mid t)$ (Appendix A, Algorithm 3). This can in turn be used to compute any marginal contrast estimand, $\tau(E[v(y)|do(t_1)], E[v(y)|do(t_2)])$.

4.2.2 Multiple Treatments and Causal Equivalence Classes

Algorithm 3 is fully general and thus can be used to compute marginal contrast estimands in a single treatment setting, multiple treatment setting, or even multiple outcome settings. However, the primary motivation in this chapter is to study this factorization in multiple treatments setting when there additional observable implications from latent variable models. The copula factorization elucidates the role of confounding even when aspects of the treatment-confounder relationship are identifiable from multiple treatments. Specifically, with multiple treatments, the conditional distribution of latent confounders given treatments, parameterized by ψ_T, is often identifiable up to a particular equivalence class (e.g. up to rotation and scale).

In this chapter, we focus on inference with latent variable methods where ψ_T is identified up to an equivalence class where the set of possible causal effects compatible with any particular value of ψ_T does not change within this class. We formalize this notion through the following definition and assumption.

Definition 4.2.1 (Causal equivalence class in multi-treatment setting) *$[\psi_T]$ is a causal equivalence class of ψ_T if and only if for any $\tilde{\psi}_T$ in $[\psi_T]$, then, for every ψ_Y there exists a $\tilde{\psi}_Y$ such that $f_{\psi_Y,\psi_T}(y \mid do(T=t)) = f_{\tilde{\psi}_Y,\tilde{\psi}_T}(y \mid do(T=t))$ for all y, t.*

For the purposes of sensitivity analysis, when ψ_T is identified up to a causal equivalence class, we can assume that ψ_T is point-identified at a particular value within the class $[\psi_T]$ without loss of generality. Identification up to a causal equivalence class is not generally possible in single treatment studies but will often hold in a multi-treatment study when certain identifying conditions for the latent variable model are met.

Crucially, the copula-based formulation enables valid sensitivity analysis without observable implications, even in these cases where ψ_T is restricted by the observed data. In this case, the outcome-confounder copula c_{ψ_Y} remains the lone degree of freedom in the sensitivity model. As we will show, this restriction can induce qualitatively different sensitivity regions in the multi-treatment setting as opposed to the single-treatment setting. For example, sensitivity regions can be bounded, even without additional restrictions on ψ_Y.

4.3 Practical Sensitivity Analysis with the Gaussian Copula

In practice, it is infeasible to characterize and interpret the implied causal effects for all possible copula specifications. In this section, we propose a practical sensitivity analysis method based on the special case in which c_{ψ_Y} is a Gaussian copula. This model characterizes the sensitivity of causal effects to monotone dependencies between the outcome and unobserved confounders. As we will discuss in the following sections, the Gaussian copula facilitates interpretation and sensitivity parameter calibration, and,

as before, is compatible with arbitrary marginal distributions $f(y \mid t)$ and $f(t)$. For our method and throughout the remainder of the dissertation we make the following additional assumptions:

Assumption 4.3.1 (Copula invariance) *The conditional copula does not depend on the value of t, that is, the conditional dependence between Y and U is invariant to the level of T.*

Assumption 4.3.2 (Gaussian copula) *The conditional copula between the outcome and m-dimensional latent confounders given treatments, $c_\psi(F_{Y|t}(y), F_{U|t}(u) \mid t)$, is a Gaussian copula.*

These assumptions do not impose constraints on the observed data distributions, only the relationship between the observed and latent variables. Given Assumption 4.3.1 and 4.3.2, the conditional confounder density can be expressed as a multivariate normal density, $f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t})$, where $\Sigma_{u|t}$ invariant to the level of t. Together, these assumptions imply the following generative model:

$$T \sim F_T \tag{4.6}$$

$$f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t}) \tag{4.7}$$

$$\tilde{Y} = \gamma'(U - \mu_{u|t}) + \epsilon_{\tilde{y}|t,u}, \quad \epsilon_{\tilde{y}|t,u} \sim N(0, \sigma^2_{\tilde{y}|t,u}), \quad \gamma^T \Sigma_{u|t} \gamma + \sigma^2_{\tilde{y}|t,u} = 1 \tag{4.8}$$

$$Y = F^{-1}_{Y|t}(\Phi(\tilde{Y})) \tag{4.9}$$

where F_T is the distribution of the treatments and $F^{-1}_{Y|t}$ is the inverse-CDF of the conditional distribution of Y given $T = t$. The Gaussian copula is parameterized by the

correlation matrix implied by

$$\text{Cov}([\tilde{Y}, U] \mid T = t) = \begin{bmatrix} 1 & \gamma^T \Sigma_{u|t} \\ \Sigma_{u|t}\gamma & \Sigma_{u|t} \end{bmatrix}. \tag{4.10}$$

with parameters are $\psi_T = \{\mu_{u|t}, \Sigma_{u|t}\}$ and $\psi_Y = \{\gamma\}$. In general, $\mu_{u|t}$ and $\Sigma_{u|t}$ will not be point identified, although under many latent variable models they can be identified up to invertible linear transformation of U. Importantly, the following theorem establishes that the class of ψ_T defined by all invertible linear transformations of U is a causal equivalance class.

Theorem 4.3.1 *Assume model 4.7-4.9. Let* $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$ *where* $\mathcal{S}^+$ *is the space of symmetric positive definite matrices. Then* $[\psi_T]$ *is a causal equivalence class.*

The gist of the proof is that for any invertible linear transformation, A, of U, the copula parameterized by $\tilde{\gamma} = A^{-1}\gamma$ yields equivalent causal effects in the reparameterized coordinates of U as γ does in the original confounder coordinates.

Throughout this chapter, we will assume that ψ_T is identified up to invertible linear transformations of U, and explore the range of possible causal effects for different γ satisfying $\gamma^T \Sigma_{u|t}\gamma \leq 1$. In Algorithm 1, we provide a procedure for estimating any marginal contrast estimand, given a sensitivity vector γ and treatments levels t_1 and t_2. At a high level, we compute a Monte Carlo estimate of $f(y \mid do(t))$ via the following three step procedure: 1) draw a sample from $f(u)$, 2) compute the conditional density of the Gaussianized outcome $f(\tilde{Y} \mid u, t)$ via the Gaussian copula and 3) transform $\tilde{Y}$ back to original space via the conditional quantile function $F^{-1}_{Y|t}$ (see Figure 4.3). In the following Sections, we introduce some theoretical insights about our approach and also provide a method for calibrating the magnitude of γ and reasoning about it's direction.

Algorithm 1: Marginal Contrast Estimation with Multiple Treatments.

Function `ComputeMean`(t, γ)**:**
  for $i = 1, 2, \ldots, n$ **do**
    $\mu_i \leftarrow \gamma^T(\mu_{u|t_i} - \mu_{u|t})$;
    for $j = 1, 2, \ldots, nSim$ **do**
      Sample $\tilde{y}_{ij}$ from $N(\mu_i, 1)$;
      $y_{ij} \leftarrow F^{-1}_{Y|t}(\Phi(\tilde{y}_{ij}))$;
  return $\frac{1}{n}\sum_{ij} v(y_{ij})$
return $\tau(\text{ComputeMean}(t_1, \gamma), \text{ComputeMean}(t_2, \gamma))$

4.4 The Geometry of Sensitivity in the Gaussian Copula Model

As described in the previous Section, our method for practical sensitivity analysis with multiple treatments is based on a Gaussian copula parameterization of the confounder-outcome relationship. Here, we start by providing some theoretical insights about how the Gaussian covariance structure relates to confounding bias in the linear-Gaussian model. Specifically, we describe how the causal effects vary as a function of γ and how bounds on these effects depend on both the treatment contrast and inferred conditional confounder density. In 4.4.2, we generalize some of these results to arbitrary models for $f(y \mid t)$ and $f(t)$.

4.4.1 Confounding Bias in the Linear-Gaussian Model

We start by illustrating our approach in a simple Linear-Gaussian model when (Y, T, U) are jointly multivariate Gaussian and establish the following results:

- For causal inference with a single treatment, the confounding bias for PATE_{t_1,t_2} is unbounded.

- When there are multiple treatments which can be used to identify (up to a causal equivalance class) the conditional confounder distribution, the magnitude of the confounding bias for PATE_{t_1,t_2} is bounded. We characterize how the magnitude of this bound depends on the parameters of the latent confounder model.

- In the multi-cause setting there are many possible treatment contrasts. The confounding bias depends on the contrast. We characterize which treatment contrasts lead to the largest bounds and which treatment contrasts imply identifiable effects.

We demonstrate these results in the following model:

$$U = \epsilon_u, \quad \epsilon_u \sim N_m(0, \Sigma_u), \tag{4.11}$$

$$T = BU + \epsilon_{t|u}, \quad \epsilon_{t|u} \sim N_k(0, \Lambda_{t|u}), \tag{4.12}$$

$$Y = \tau' T + \gamma' U + \epsilon_{y|t,u}, \quad \epsilon_{y|t,u} \sim N(0, \sigma^2_{y|t,u}), \tag{4.13}$$

with $\tau \in \mathbb{R}^k$, $\gamma \in \mathbb{R}^m$, and $\Lambda_{t|u}$ an arbitrary diagonal matrix. When either $B = 0$ or $\gamma = 0$, there is no confounding. We also note that Equations 4.11 and 4.12 imply that the conditional distribution of the confounder can be expressed as $f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t})$, as in Equation 4.7, where both $\mu_{u|t}$ and $\Sigma_{u|t}$ are known functions of B and $\sigma^2_{t|u}$. Under model 4.11-4.13, the intervention distribution has density

$$f(y \mid do(T = t)) \sim N(\tau' t,\ \sigma^2_{y|t,u} + \gamma' \Sigma_u \gamma). \tag{4.14}$$

For any t_1, t_2, PATE_{t_1,t_2} is characterized entirely by the regression coefficients τ. The observed outcome distribution can be expressed as

$$f(y \mid T = t) \sim N(\tau'_{\text{naive}} t,\ \sigma^2_{y|t}), \tag{4.15}$$

where

$$\tau_{\text{naive}} = \tau + (B\Sigma_u B' + \Lambda_{t|u})^{-1} B\Sigma_u \gamma \tag{4.16}$$

$$\sigma^2_{y|t} = \sigma^2_{y|t,u} + \gamma'(\Sigma_u - \Sigma_u B'(B\Sigma_u B' + \Lambda_{t|u})^{-1} B\Sigma_u)\gamma \tag{4.17}$$

$$= \sigma^2_{y|t,u} + \gamma' \Sigma_{u|t} \gamma \tag{4.18}$$

which are both fully identified from observed data. We refer to τ_{naive} as the naive estimate since it naively neglects the effect of unobserved confounders. Equation 4.18 shows that the observed residual outcome variance can be decomposed into nonconfounding variation $\sigma^2_{y|t,u}$ and confounding variation, $\gamma' \Sigma_{u|t} \gamma$. We take $\sigma^2_{y|t}$ and τ_{naive} as fixed and known, and characterize the range of confounding biases by considering different assumptions about the strength of confounding.

We note that the bias of the naive estimator depends only on the difference between the treatment vectors, $\Delta t = t_1 - t_2$, since the population average treatment effect can be expressed as

$$\text{PATE}_{\Delta t} = \tau'(t_1 - t_2) := \tau' \Delta t, \tag{4.19}$$

The confounding bias, denoted $\text{Bias}_{\Delta t} = \tau'_{\text{naive}} \Delta t - \text{PATE}_{\Delta t}$, can then be expressed as

$$\text{Bias}_{\Delta t} = \gamma' \Sigma_u B' (B\Sigma_u B' + \Lambda_{t|u})^{-1} \Delta t \tag{4.20}$$

$$= \gamma'(E(U \mid T = t_1) - E(U \mid T = t_2)) \tag{4.21}$$

$$:= \gamma' \mu_{u|\Delta t}, \tag{4.22}$$

where we use $\mu_{u|\Delta t}$ to denote the difference in confounder means for the treatment contrast, Δt. We can then succinctly express the PATE in terms of the naive estimate minus the bias as $\text{PATE}_{\Delta t} = \tau_{\text{naive}}{}' \Delta t - \gamma' \mu_{u|\Delta t}$.

In the single treatment setting, neither $\psi_Y = \{\gamma\}$ nor $\psi_T = \{\mu_{u|\Delta t}, \Sigma_{u|t}\}$ are identifiable, which implies that the confounding bias is unbounded. However, with multiple treatments ψ_T is identifiable up to a causal equivalence class defined by invertible linear transformations of U. We make this concrete in the following theorems.

Theorem 4.4.1 *Suppose that the observed data is generated by model 4.11-4.13. When there k treatments with $1 < m < k$, then ψ_T is identified up to the causal equivalence class $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$. When there is a single treatment ($k = 1$) or at least $m = k$ confounders, then ψ_T is not identifiable up to causal equivalence class.*
Proof. See appendix.

One consequence of Theorem 4.4.1 is that the distribution of U is only causally relevant up to linear transforms, and as such, without loss of generality, we make the simplifying assumption that $U \sim N(0, I_m)$ for the remainder of this Section.

First, we review the implications of this theorem when there is only a single treatment, i.e. $k = 1$. As shown in Cinelli and Hazlett (2019) [18], for single treatment inference, the squared confounding bias of the PATE can be expressed as

$$Bias^2_{\Delta t} = \frac{R^2_{T\sim U}}{1 - R^2_{T\sim U}} R^2_{Y\sim U|T} \frac{\sigma^2_{y|t}}{\sigma^2_T} \tag{4.23}$$

where $\sigma^2_T := BB' + \sigma^2_{t|u} I_k$ is the marginal variance of the treatment,

$$0 \leq R^2_{T\sim U} = \frac{\sigma^2_T (\mu_{u|\Delta t})' \mu_{u|\Delta t}}{(\Delta t)^2} \leq 1 \tag{4.24}$$

is the unidentifed fraction of treatment variance explained by confounders and

$$0 \leq R^2_{Y\sim U|T} = \frac{\gamma^T \Sigma_{u|t} \gamma}{\sigma^2_{y|t}} \leq 1 \tag{4.25}$$

is the fraction of the residual outcome variance explained by confounders. By Theorem 4.4.1, neither $R^2_{T\sim U}$ nor $R^2_{Y\sim U|T}$ are identifiable. Since $\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}$ can be arbitrarily large, the confounding bias is unbounded in the single treatment setting.

In contrast, Theorem 4.4.1 states that with multiple treatments, we can identify an element of the causal equivalence class for parameters governing the conditional confounder distribution. The relationship between Y and U, as parameterized by m-vector γ, remains an unidentified sensitivity vector. This sensitivity vector can be viewed as parameterizing the Gaussian conditional copula between Y and U given T, $c_\gamma(F_{Y|t}(y), F_{U|t}(u) \mid t)$ (Equation 4.10).

Identification up to causal equivalence class implies that the confounding bias is bounded. From Equation 4.20, we can see that the sign and magnitude of the bias depends on both Δt as well as γ. Although γ is not identified, its values are constrained since unobserved confounding cannot explain more than 100% of the residual outcome variance (Equation 4.25). This constraint on the magnitude of γ implies the following result about the bias of the naive estimator.

Theorem 4.4.2 *Suppose that the observed data is generated by model 4.11-4.13 with* $\sigma^2_{t|u} > 0$. *Then,* $\forall\gamma$ *satisfying Assumptions 1 and 2,*

$$\gamma^T \Sigma_{u|t} \gamma \leq \sigma^2_{y|t} \tag{4.26}$$

For any given Δt, *we have*

$$Bias^2_{\Delta t} \leq \sigma^2_{y|t} R^2_{Y\sim U|T} \|\Sigma^{-1/2}_{u|t} \mu_{u|\Delta t}\|^2_2. \tag{4.27}$$

The bound is achieved when γ *is colinear with* $\Sigma^{-1}_{u|t}\mu_{u|\Delta t}$.

Proof. See appendix.

This theorem states that the true causal effect lies in the interval $\tau'_{naive}\Delta t \pm \sqrt{\sigma^2_{y|t} R^2_{Y\sim U|T}} \|\Sigma^{-1/2}_{u|t}\mu_{u|\Delta t}\|_2$. When additional assumptions are applied to identify a particular value of γ (e.g. see [45]), the corresponding causal effect estimate will correspond to a single point inside this ignorance region when the Gaussian copula assumption holds. We refer to the right-hand side of 4.27 as the "worst-case bias" of the naive estimator. In particular, since τ_{naive} is the midpoint of the ignorance region, it has the minimum worst-case bias over all alternative causal effect estimators. This is consistent with Grimmer et al. (2020) [41] who emphasize that the deconfounder proposed by Wang and Blei (2018) [36] cannot outperform the naive estimator in general.

The worst-case bias of τ_{naive} is proportional to the norm of the scaled difference in confounder means in each treatment arm, $\Sigma^{-1/2}_{u|t}\mu_{u|\Delta t}$. This result provides a useful generalization of existing work which demonstrates that overlap is violated when u can be pinpointed by a deterministic function of t [38]. In contrast to the original work by Blei (2018) [36], our result suggests that the more precisely we can pinpoint u given t, the *less* precisely we can pinpoint $\text{PATE}_{\Delta t}$. In the following corollary, we assume $\Lambda_{t|u} = \sigma^2_{t|u} I_k$ for gaining intuitions about the worst-case bias over all possible treatment contrasts:

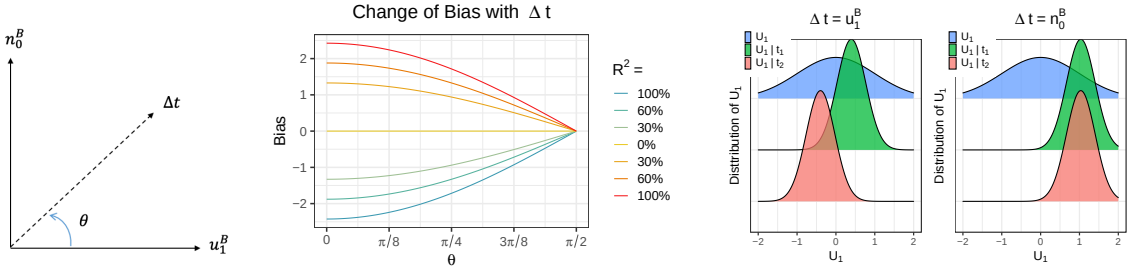
Corollary 4.4.1 *Let d_1 be the largest singular value of B. For all Δt with $\| \Delta t \|_2 = 1$, the squared bias is bounded by*

$$Bias^2_{\Delta t} \leq \frac{d_1^2}{(d_1^2 + \sigma^2_{t|u})} \frac{\sigma^2_{y|t}}{\sigma^2_{t|u}} R^2_{Y\sim U|T}, \tag{4.28}$$

with equality when $\Delta t = u_1^B$, the first left singular vector of B. When $\Delta t \in Null(B')$, the naive estimate is unbiased, that is, $PATE_{\Delta t} = \tau'_{naive}\Delta t$.

Proof: See Appendix.

The first term in (4.28), $\frac{d_1^2}{(d_1^2+\sigma^2_{t|u})}$, is the fraction of variance in the first principal component of the causes that can be explained by confounding. The first principal component



(a) $\Delta t := t_1 - t_2$ (b) Estimation Bias of $\text{PATE}_{\Delta t}$ (c) Distribution of U_1

Figure 4.2: Illustration of Corollary 4.4.1. (a) We parameterize Δt with θ, the angle between n_0^B, a vector in the null space of B, and u_1^B, the first left singular vector of B. (b) The confounding bias of naive estimates of $\text{PATE}_{\Delta t}$ changes with θ and depends on $R^2_{Y\sim U|T}$. (c) Confounder densities in different populations. The blue, green, red densities denote distributions of U_1 in the observed population, the subpopulation receiving t_1 and the subpopulation receiving treatment t_2 respectively. Observed data estimates of $\text{PATE}_{\Delta t}$ are unbiased when $\Delta t = n_0^B$, since the confounder distributions are the same in two treatment arms. However, observed data estimates of $\text{PATE}_{t_1,\bullet}$ and $\text{PATE}_{t_2,\bullet}$ are biased since in general the superpopulation distribution of the confounder is different.

corresponds to the projection of treatments which is most correlated with confounders, and thus is the causal contrast with the largest ignorance region. We also note that the squared biases depends on $R^2_{Y\sim U|T}$, the partial variance explained by confounders given treatments. While the magnitude of the confounding bias is always largest when $R^2_{Y\sim U|T} = 1$, we often have reason to believe that the variance explained by confounders is likely smaller. We describe how to leverage this idea to calibrate more plausible bounds and measures of robustness in Section 4.5.

We illustrate some key insights from these theorems in Figure 4.2, where we display the worst-case bias as a function of the treatment contrasts, Δt. In this illustration, we assume that Δt lies on a plane spanned by u_1^B and n_0^B, an arbitrary vector in the null space of B. We let $\theta = \arccos(\Delta t' u_1^B)$ be the angle of Δt relative to u_1^B, Figure 4.2a. Figure 4.2b depicts the bias as function of θ for different values of $R^2_{Y\sim U|T}$. When Δt is in the null space of B, $\text{PATE}_{\Delta t}$ is identified because the confounder distributions are identical in the two treatment arms, i.e. there is no confounding for this particular

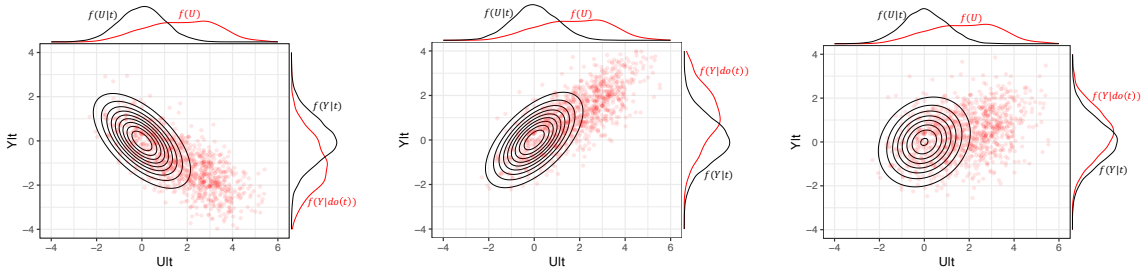
contrast. When Δt is colinear with u_1^B the scaled difference in means of u is largest, which implies the largest worst-case bias for the treatment effect, Figure 4.2c (left). Even when $\text{PATE}_{\Delta t}$ is identified, we emphasize that $\text{PATE}_{t_1,\bullet}$ and $\text{PATE}_{t_2,\bullet}$ are both biased, since the distribution of confounders in the treatment arm differs from the distribution of confounder in the superpopulation, Figure 4.2c (right). As noted by others, identification of $\text{PATE}_{\Delta t}$ for Δt in the null space of B arises due to bias cancellation in intervention means of the two treatment arms [41].

Our theory is invariant to rotations of the treatments vector, which means that under model 4.11-4.13, we can always make a change of treatment variables so that each confounder affects a distinct single rotated treatment (called "single cause" confounders in Wang and Blei (2018) [36]). Specifically, let $\tilde{T} = RT$ be a rotation of the original treatment variables, so that $\tilde{T} \sim N(0, \Delta + \sigma^2_{t|u} I_k)$ where Δ is a diagonal positive semi-definite matrix. Then, as before, $\frac{\Delta_i}{\Delta_i + \sigma^2_{t|u}}$ is the fraction of variance in the ith rotated treatment due to confounding and provides a bound on the omitted confounder bias of estimates of PATE for $\Delta\tilde{T} = e_i$. For the $k - m$ contrasts corresponding to the zeros in the diagonal of Δ, there is no confounding for these particular treatment contrasts and thus there is no confounding bias; these correspond to the contrasts in the original space which fall in the null space of B.

4.4.2 Generalizing the Linear-Gaussian Model

Next, we generalize beyond the setting in which (Y, T, U) is jointly multivariate Gaussian. First, when Y is Gaussian with mean $\mu_{y|t}$ and variance $\sigma^2_{y|t}$, even when T has an arbitrary functional relationship with Y we have that

$$\text{PATE}_{t,\bullet} = (\mu_{y|t} - \mu_y) - \sigma_{y|t}\gamma'(\mu_{u|t} - \mu_u), \tag{4.29}$$



(a) $\sqrt{R^2_{Y\sim U|T}} = 0.6$ (b) $\sqrt{R^2_{Y\sim U|T}} = 0.6$ (opposite sign) (c) $\sqrt{R^2_{Y\sim U|T}} = 0.2$

Figure 4.3: Differences between observed and intervention densities as a function of the fraction of outcome variance explained by a single confounder. The black contours depict the conditional Gaussian copula, $c_\gamma(F_{Y|t}(y), F_{U|t}(u) \mid t)$ whereas red points represent samples from the joint distribution, $f(y, u \mid do(t)) \propto f(y \mid t)c_\gamma(F_{Y|t}(y), F_{U|t}(u) \mid t)f(u)$. We visualize the shift in the outcome density for different conditional correlations and note that smaller values of $R^2_{Y\sim U|T}$ imply smaller biases in the outcome despite large imbalance in the distribution of U.

where $\mu_u = E[U]$ is the population mean of U. Thus,

$$\text{PATE}_{t_1,t_2} = (\mu_{y|t_1} - \mu_{y|t_2}) - \sigma_{y|t}\gamma'(\mu_{u|t_1} - \mu_{u|t_2}). \tag{4.30}$$

This leads to the following generalization of Theorem 4.4.2.

Theorem 4.4.3 *Assume the model 4.7-4.9 with Gaussian outcomes. If $\Sigma_{u|t}$ is non-invertible, then $Bias_{t_1,t_2}$ is bounded if and only if $\mu_{u|t_1} - \mu_{u|t_2}$ is in the row space of $\Sigma_{u|t}$. When bounded,*

$Bias^2_{t_1,t_2} \leq \sigma^2_{y|t} R^2_{Y\sim U|T} \|(\Sigma^\dagger_{u|t})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2})\|_2^2$, where $\Sigma^\dagger_{u|t}$ is the pseudo-inverse of $\Sigma_{u|t}$.

Proof: See Appendix.

As before, when bounded, the bias is proportional to the norm of the scaled difference in confounder means in the two treatment arms.

When there exists an m-vector, q, such that $\text{Var}(q'U \mid t) = 0$, then $\Sigma_{u|t}$ is non-invertible because there exists a projection of the confounders which is point identified.

Theorem 4.4.3 says that in this case, the ignorance region for the PATE is bounded if and only if $q'(\mu_{u|t_1} - \mu_{u|t_2}) = 0$. In words, if a projection of the confounders can be identified, then the confounding bias is bounded if the identifiable projection of the confounders has the same value in both treatment arms. This result can be viewed in the context of Theorem 7 in Wang and Blei (2018) [36], which assumes consistency and overlap of estimators for the relevant latent confounders for identification.

When the observed outcome distribution is non-Gaussian, we cannot necessarily express PATE_{t_1,t_2} analytically. In particular, for non-Gaussian Y, when $f(u \mid t_1) \sim f(u \mid t_2)$ the average treatment effect among the t_1 and t_2 treated units is unconfounded, but the bias of PATE_{t_1,t_2} may be nonzero since $f(u \mid t) \nsim f(u)$. The causal effects, however, can still be calculated using Algorithm 1. Another particularly important non-Gaussian setting we highlight here is when the outcome is binary. Interestingly, unlike the linear case, $RR_{t,\bullet}$ and RR_{t_1,t_2} are non-monotone in the magnitude of γ. We discuss this in more detail in Appendix A.2.3 and provide simulation results with binary outcomes in Section 4.6.

4.5 Calibration and Robustness

Sensitivity analyses consist of two parts: first, the sensitivity model itself, which specifies a set of causal models, indexed by sensitivity parameters; and secondly, exploratory tools for mapping external assumptions to particular causal models in this set. We now turn to discussing the latter in the context of our proposed model.

In the sensitivity analysis literature so far, two exploratory techniques have been particularly popular in single treatment studies: *calibration*, which maps sensitivity parameter values to interpretable observable or hypothetical quantities; and *robustness analysis*, which characterizes the "strength" of confounding necessary to change the con-

clusion of a study. Here, we show how to adapt these techniques to our sensitivity model in the multi-treatment setting. In addition, we introduce a third class of tools that are particularly well-suited to the multi-treatment setting, which we call *multiple contrast criteria* (MCCs). MCCs specify aggregate properties of the treatment effects for multiple treatment contrasts that are implied by a single causal model, e.g., the L2 norm of PATEs corresponding to contrasts in each individual treatment variable in T. In many multi-treatment settings, assumptions are often expressed in terms of the aggregates—e.g., in genomics, the idea that the effect of most single nucleotide polymorphisms is small—and we show here how these can be used in conjunction with our sensitivity model to characterize candidate causal models that may be of interest in an application.

4.5.1 Calibration for a Single Contrast

We begin by describing calibration for γ in our sensitivity model when the focus is on a single treatment contrast, between levels $T = t_1$ and $T = t_2$. The goal is to develop heuristics for specifying "reasonable" values or ranges for γ, e.g., to derive bounds on treatment effects by specifying bounds on the strength or direction of confounding. Following previous work in the single treatment setting, we outline how to calibrate our sensitivity parameter vector γ in terms of a fraction of outcome variance explained by the unobserved confounder. Recall that γ is a vector that parameterizes the residual correlation between the m-dimensional unobserved confounder U and the outcome Y after conditioning on the treatment vector T.

First, we briefly review calibration in single-treatment settings. In latent variable approaches for single treatment sensitivity analysis, the causal effect is identified given two sensitivity parameters: the fraction of outcome variance explained by unobserved confounders, $R^2_{Y\sim U|T}$, and the fraction of treatment variance explained by unobserved

confounders, $R^2_{T\sim U}$ [18]. In a linear model, these two scalar quantities identify the confounding bias (Equation 4.23). Neither R-squared value is identifiable and thus many authors have proposed strategies for drawing analogies between these values and other observable or hypothetical quantities [48, 19, 2].

We borrow this strategy for calibration in our setting, with some modifications. First, in our setting there is no need to calibrate $R^2_{T\sim U}$, because we have restricted ourselves to the case where $f(u \mid t)$ is identified up to a causal equivalence class. This leaves calibration of the outcome-confounder relationship, which in our setting is more complex because it is parameterized by a vector γ[1]. However, we can reparameterize γ in terms of a direction d and an R-squared for interpretable calibration:

$$\gamma = \sqrt{R^2_{Y\sim U|T}}\Sigma^{-1/2}_{u|t}d, \tag{4.31}$$

where $d \in \mathbb{S}^{m-1}$ is an m-dimensional unit vector. We discuss strategies for calibrating both the magnitude and direction separately.

Calibrating the magnitude of γ. For Gaussian outcomes, the magnitude of γ is characterized entirely by $R^2_{Y\sim U|T}$, the partial fraction of outcome variance explained by U given T. When $R^2_{Y\sim U|T} = 0$ there is no unobserved confounding and when $R^2_{Y\sim U|T} = 1$, all of the observed residual variance in Y is due to confounding factors. In order to calibrate this magnitude, we adopt an idea proposed by Cinelli and Hazlett (2019) [18] for inference with single treatments. In their work, they calibrate $R^2_{Y\sim U|T}$ by comparing it to the partial fraction of variance explained by different observed covariates. We use a closely related strategy that makes use of the presence of multiple treatments rather than observed covariates. Specifically, we compute the fraction of variation in Y that can be explained by a specific treatment (or set of treatments), T_j, after controlling for

[1]Unlike the single treatment setting, the confounder-outcome relationship cannot be sufficiently summarized in terms of a scalar $R^2_{Y\sim U|T}$. Each confounder can impact each treatment in different ways.

all other treatments T_{-j} as

$$R^2_{Y\sim T_j|T_{-j}} := \frac{R^2_{Y\sim T} - R^2_{Y\sim T_{-j}}}{1 - R^2_{Y\sim T_{-j}}}. \tag{4.32}$$

When observed covariates are available, we can still analogously compute the partial fraction of variance explained by an observed covariate, $R^2_{Y\sim X_j|T,X_{-j}}$, as done in Cinelli and Hazlett (2019) [18]. Veitch and Zaveri (2020) [19] and Cinelli et al. (2020) [48] propose useful graphical summaries for calibration based on these metrics in the single treatment setting.

When the observed outcome is non-Gaussian, we calibrate the "implicit R^2", by considering the explained variance of the latent Gaussian outcome, $\tilde{Y}$ in Equation 4.8. The implicit R^2 of T for model 4.7-4.9 is defined as

$$R^2_{\tilde{Y}\sim T} = \frac{\text{Var}(E[\tilde{Y}|T])}{\text{Var}(E[\tilde{Y}|T]) + 1}. \tag{4.33}$$

and the implicit partial R-squared of treatment T_j, $R^2_{\tilde{Y}\sim T_j|T_{-j}}$, is defined analogously to Equation 4.32. As before, these estimable partial R-squared values can be used to provide a useful comparison for the partial R-squared of potential unobserved confounders, $R^2_{\tilde{Y}\sim U|T}$. For more detail, see Imbens (2003) [3] and Franks et al. (2019) [2] who discuss calibration with implicit R-squared values in logistic regression models.

Choosing the direction of γ. Given a magnitude, we now propose a default method for identifying the direction of γ for a single contrast. The dot product $d'\Sigma^{-1/2}_{u|t}(\mu_{u|t_1} - \mu_{u|t_2})$ corresponds to the projection of the scaled difference in confounder means onto the outcome space. By default, we suggest using the direction which maximizes the squared bias. As shown in Theorem 4.4.3, when d is colinear with $\Sigma^{-1/2}_{u|t}(\mu_{u|t_1} - \mu_{u|t_2})$,

the confounding bias of the naive estimator for Gaussian outcomes is maximized at

$$|\text{Bias}_{t_1,t_2}| = \sqrt{R^2_{Y\sim U|T}}\sigma_{y|t}\|\Sigma^{-1/2}_{u|t}(\mu_{u|t_1} - \mu_{u|t_2})\|_2, \tag{4.34}$$

Choosing the direction of the sensitivity vector in this way provides conservative bounds for each contrast of interest. For non-Gaussian outcomes or alternative estimands, there may not be an analytic solution to the direction which maximizes the bias, but we can still compute the direction via numerical optimization.

4.5.2 Robustness for Single Contrasts

We now turn to assessing the robustness of conclusions using our sensitivity model, extending work by Cinelli and Hazlett (2019) [18] and VanderWeele and Ding (2017) [49] in the single treatment setting. Specifically, we propose an extension of the robustness value (RV), which characterizes the minimum strength of confounding needed to change the sign of the treatment effect. As in the previous section, the extension is most straightforward when considering the effect of a single treatment contrast, between levels $T = t_1$ and $T = t_2$.

To review briefly, in single treatment settings, Cinelli and Hazlett define the RV as the minimum R-squared needed to reduce the treatment effect to zero, assuming $R^2_{Y\sim U,T} = R^2_{T\sim U}$. (we return to this assumption below.) A robustness value close to one means the treatment effect maintains the same sign even if nearly all of the observed residual variance in the outcome is due to confounding. On the other hand, a robustness value close to zero means that even weak confounding would change the sign of the point estimate.

In the multi-treatment setting, we can more precisely characterize the robustness of causal effects when $R^2_{T\sim U}$ is identifiable. In particular, we can compute an analogue to the

RV without assuming $R^2_{Y\sim U,T} = R^2_{T\sim U}$, an assumption that, in single treatment settings, can be consequential[2]. With $R^2_{T\sim U}$ known, we define multi-treatment RV can then simply as the minimum value of $R^2_{Y\sim U|T}$ needed to explain away the treatment effect of interest, assuming the direction of the sensitivity vector is chosen to maximize the bias. When the observed outcomes are Gaussian, the robustness value can be computed in closed form:

$$\mathrm{RV}_{t_1,t_2} = \frac{(\mu_{y|t_1} - \mu_{y|t_2})^2}{\sigma^2_{y|t}\|\Sigma^{-1/2}_{u|t}(\mu_{u|t_1} - \mu_{u|t_2})\|^2}, \tag{4.35}$$

RV metrics for alternative estimands and/or non-Gaussian data can still be computed using the same principle. For example, when the observed outcome is binary, the RV can be computed numerically by solving $RR_{t_1,t_2} = 1$, which corresponds to the minimum strength of confounding needed for the observed risk ratio (RR) to equal to one. This robustness value is analogous to the "E-value" proposed by VanderWeele and Ding (2017) [49].

In our setting, we can also make stronger statements about robustness than in the single treatment setting: under the latent variable model, it is possible to declare an effect robust to *any* level of confounding. In particular, when the latent variable model implies $R^2_{T\sim U} < 1$ (i.e., we have confounder overlap), then even when $R^2_{Y\sim U|T} = 1$, the ignorance region is bounded (Theorem 4.4.3). When this ignorance region excludes zero, we declare the effect "robust". This operation is consistent with the result in Miao et al. (2018) [50], showing that hypotheses of zero effect can be tested in this setting, even if the treatment effect cannot be identified.

[2]In single treatment settings, when $R^2_{Y\sim U|T} > R^2_{T\sim U}$, the single-treatment RV will be too conservative. Conversely, when $R^2_{Y\sim U|T} < R^2_{T\sim U}$ the single-treatment RV will overestimate the robustness of the effect.

4.5.3 Calibration for a Single Contrast with Null Controls

We also find it interesting to include additional constraints into our sensitivity analysis, especially null control treatments, which are powerful tools to detect and adjust for bias in many application, such as Genome Wide Association Study (GWAS) and epidemiological research [51]. A set of null control treatments would be subset of treatments that were known a priori to have zero (or bounded) causal effect on the outcome, for example genes in a GWAS which were known to be causally unrelated to a particular phenotype.

Much of the progress that has been made in removing unobserved confounding, focusing on inference about identification, with null controls under relatively strong assumptions. [52, 53, 54, 55, 56, 57, 58] use paternal exposure as the null control treatment to study the intrauterine effect of maternal exposure on offspring outcome. [59, 60, 61, 62] utilize future air pollution as the null control treatment to detect and reduce the confounding bias of the estimated effect of air pollution on diseases. Here, we illustrate that, with relaxed assumptions when a set of null controls is insufficient to identify causal effects, they still reduce the ignorance regions by imposing additional constraints on confoundedness. We demonstrate the idea especially in the Gaussian outcome cases.

Let $\mathcal{C}$ be a set indexing c null control treatments contrasts, t_{j1} and t_{j2}, such that for any $j \in \mathcal{C}$, we have $\text{PATE}_{t_{j1},t_{j2}} = 0$. For these null controls, the *observed* mean difference in outcomes, defined as a row vector $\mu_{y|\Delta t_c} := [(\mu_{y|t_{11}} - \mu_{y|t_{12}}), \cdots, (\mu_{y|t_{c1}} - \mu_{y|t_{c2}})]$, must equal the omitted confounder bias. Since the bias is a function of the sensitivity vector, γ, we can establish a constraint on γ via the equation

$$\sigma_{y|t}\gamma'\Sigma_{u|t}^{1/2}M_{u|\Delta t_{\mathcal{C}}} = \mu_{y|\Delta t_c} \tag{4.36}$$

where $M_{u|\Delta t_{\mathcal{C}}} = (\Sigma_{u|t}^{\dagger})^{1/2}[(\mu_{u|t_{11}} - \mu_{u|t_{12}}), \cdots, (\mu_{u|t_{c1}} - \mu_{u|t_{c2}})]$ is a $m \times c$ matrix with

columns corresponding to the scaled difference in confounder means for each null control. Equation 4.36 has two important implications. First, we have a testable constraint that $\mu_{y|\Delta t_c}$ must be in the row space of $M_{u|\Delta t_{\mathcal{C}}}$ in order for the null control assumption to be valid. Second, in order to make up the difference between the observed effect and the true null effect of the control treatments, the magnitude of γ must be large enough. That is, constraint 4.36 implies a lower bound on the fraction of variance explained by confounders. We formalize these ideas in the following theorem.

Proposition 4.5.1 *Suppose there are c known null control treatment contrasts, t_{j1} versus t_{j2} for $j \in \mathcal{C}$. Then, the null control compatibility condition $\mu_{y|\Delta t_c} P_{M_{u|\Delta t_c}} = \mu_{y|\Delta t_c}$ must hold where $P_{M_{u|\Delta t_c}}$ denotes the projection matrix into the row space of $M_{u|\Delta t_c}$. Additionally, the partial fraction of variance explained due to confounders given treatments is lower bounded by*

$$R^2_{Y\sim U|T} \geq R^2_{min} = \frac{1}{\sigma^2_{y|t}} \parallel \mu_{y|\Delta t_c} M^{\dagger}_{u|\Delta t_{\mathcal{C}}} \parallel^2_2 .$$

where $M^{\dagger}_{u|\Delta t_{\mathcal{C}}}$ denotes a generalized inverse of $M_{u|\Delta t_{\mathcal{C}}}$.
Proof: See Appendix.

Although the null controls assumption implies a *lower bound* on the magnitude of confounding, null controls actually reduce the width of the partial identification region, which is no longer centered at the naive estimate of the treatment effect. We characterize the partial identification region under the null controls assumption in the following Theorem.

Theorem 4.5.1 *For any value of $R^2_{Y\sim U|T} > R^2_{min}$ which satisfies null control compatibility condition, the confounding bias for the treatment effect of contrast Δt is in the*

interval

$$\mu_{y|\Delta t_c} M^{\dagger}_{u|\Delta t_{\mathcal{C}}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \pm \tag{4.37}$$

$$\sigma_{y|t} \sqrt{R^2_{Y \sim U|T} - R^2_{\min}} \left\| Q^{\perp}_{M_{u|\Delta t_c}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \right\|_2 \tag{4.38}$$

where $Q^{\perp}_{M_{u|\Delta t_c}}$ *is the* $m \times m$ *projection matrix into the complement of the column space of* $M_{u|\Delta t_c}$.

From Theorem 4.5.1 the following corollary follows immediately.

Corollary 4.5.1 *Under the assumptions established in Theorem 1, null controls reduce the width of the partial identification ignorance region by a multiplicative factor of*

$$\sqrt{1 - R^2_{min}/R^2_{Y \sim U|T}} \frac{\| \, Q^{\perp}_{M_{u|\Delta t_c}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \, \|_2}{\| \, (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \, \|_2} \leq 1 \tag{4.39}$$

This corollary highlights that there are two ways in which null control reduce the width of the worst-case ignorance region: the first term under the radical shows that null controls constrain the magnitude of the confounding bias (Proposition 4.5.1) which proportionally reduces the width of the ignorance region for all contrasts by an equal amount. The second term is contrast dependent and indicates that null controls reduce the ignorance the most for treatment contrasts that have the most similar confounder mean differences. For a treatment contrast Δt, when $(\mu_{u|t_1} - \mu_{u|t_2})$ is in the span of the row space of $M_{u|\Delta t_c}$, then the treatment effect is identified; when $(\mu_{u|t_1} - \mu_{u|t_2})$ is orthogonal to the row space of $M_{u|\Delta t_c}$ then there is no further reduction in the ignorance region for PATE, beyond the constraint on the magnitude. In summary, the best null controls are those which have large confounding bias and also have confounder distributions similar to those in the treatment contrasts of interest. A direct consequence of Theorem 4.5.1 is that when $M_{u|\Delta t_{\mathcal{C}}}$ is full rank, *all* treatments are identifiable.

4.5.4 Multiple Contrast Criteria

In addition to exploratory tools used with single-treatment sensitivity analysis, the multi-treatment setting presents opportunities for exploring sensitivity models in new ways. Here, we propose one such approach using multiple contrast criteria, or MCCs. As opposed to the approaches we have discussed so far, which consider treatment contrasts in isolation, MCCs characterize a choice of sensitivity vector γ by concurrently considering its implications for multiple treatment contrasts in aggregate. Thus, while the sensitivity vector γ that gives the worst-case bias may differ across individual contrasts, an MCC can be used to select a single γ that has implications for many simultaneous treatment effects.

Formally, for a set of treatment contrasts $\mathcal{T}^2 = \{(t_1, t_2)_k\}_{k=1}^K$, and a candidate sensitivity vector γ, let $\mathbf{PATE}_{\mathcal{T}^2}(\gamma)$ be the vector of PATEs implied by the causal model indexed by γ. An MCC is a scalar summary of this treatment effect vector, which we write as $\omega(\mathbf{PATE}_{\mathcal{T}^2}(\gamma))$. An MCC is specified by the set of constrasts $\mathcal{T}^2$ and the summary function ω, both of which can be chosen to meet the needs of a given analysis.

MCCs can be used in many ways, but here we consider how they can be used to search for the causal model that yields the minimum norm treatment effect vector, subject to assumptions 1.2.1-4.3.2 and a confounding limit $\mathcal{R}^2$. Specifically, we take ω to be an L_p norm for some p, and consider sensitivity vectors γ_* that satisfy:

$$\gamma_* = \underset{\tilde{\gamma}}{\text{argmin}}\ \omega(\mathbf{PATE}_{\mathcal{T}^2}(\tilde{\gamma})) \text{ subject to } R^2_{Y\sim U|T}(\tilde{\gamma}) \leq \mathcal{R}^2 \tag{4.40}$$

where $R^2_{Y\sim U|T,X}(\tilde{\gamma}) = \frac{\tilde{\gamma}'\Sigma_{u|t}\tilde{\gamma}}{\sigma^2_{y|t}}$ is the partial fraction of outcome variance explained by confounding for sensitivity vector $\tilde{\gamma}$. Causal models selected in this way are often highly interpretable, in terms of either "worst case" effect sizes or established prior knowledge. For example, we can chose ω to be the L_∞ norm, so that γ_* is the sensitivity vector

that minimizes the maximum absolute effect across contrasts. Alternatively, we could choose ω to be the L_1 or L_2 norm of the treatment effects to find models that imply small "typical" effect sizes. We demonstrate how this minimization approach can be used to express prior knowledge about small effects in simulated data in Section 4.6.2, and how it can be used to evaluate robustness on a real data set in Section 4.7.

4.6 Simulation Studies

In this Section, we demonstrate our sensitivity analysis workflow in several numerical simulations. The goal of these simulations is twofold: first, to demonstrate some of the operating characteristics of the approach in settings that are more realistic than the linear Gaussian settings we characterized analytically; and secondly, to show how exploratory tools like calibration, robustness analysis, and MCCs can be used to draw conclusions and choose interesting candidate models.

We consider two broad simulation settings. In the first setting, we construct simulations with non-linear responses to treatment to show how the ignorance regions returned by our method can vary in different scenarios. In the second setting, we construct a simulation that mimics the structure of a Genome Wide Association Study (GWAS). Here, we examine the behavior of our method when a popular approximate latent variable method—the Variational Auto Encoder (VAE)—is used to estimate the effects of latent confounders, and demonstrate how MCCs can be useful tools for using prior information to choose potentially useful causal models from the set that is compatible with the observed data. In both subsections, we simulate data from the following generating

process:

$$U := \epsilon_u, \quad \epsilon_u \sim N_m(0, I_m), \tag{4.41}$$

$$\tilde{T} := BU + \epsilon_{t|u}, \quad \epsilon_{t|u} \sim N_k(0, \sigma^2_{t|u} I_k), \tag{4.42}$$

$$T := h_T(\tilde{T}), \tag{4.43}$$

$$\tilde{Y} := g(T) + \gamma' U + \epsilon_{y|t,u}, \quad \epsilon_{y|t,u} \sim N(0, \sigma^2_{y|t,u}), \tag{4.44}$$

$$Y := h_{Y|T}(\tilde{Y}), \tag{4.45}$$

The functions h_Y and h_T are chosen according to be either the identity for Gaussian data, or an indicator function for binary data.

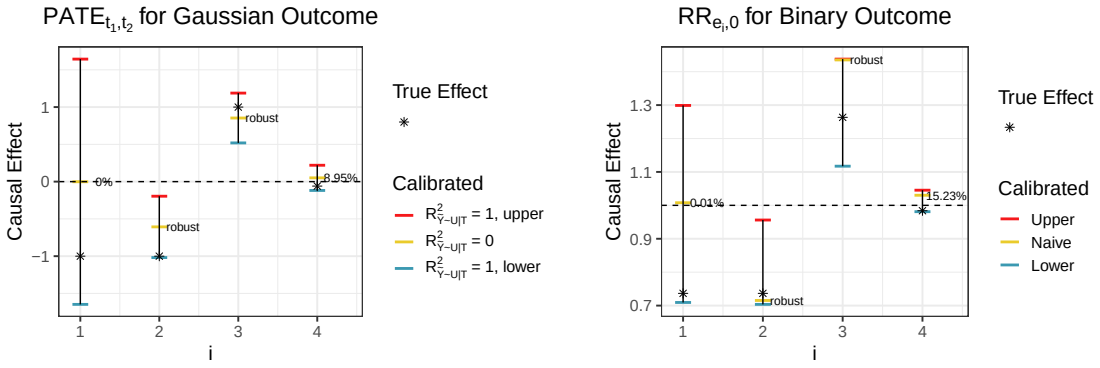
4.6.1 Example with Non-Linear Response Functions

We start by exploring variation in the size of ignorance regions for different contrasts in a simple simulated example with four treatments where the outcome is a nonlinear function of these treatments. We consider two cases: first, a case where Y is Gaussian with $h_Y(\tilde{Y}) = \tilde{Y}$; and secondly, a case where Y is binary with $h_Y(\tilde{Y}) = I_{\tilde{Y}>0}$. We aim to estimate the $\text{PATE}_{e_i,0}$ for Gaussian outcome and $\text{RR}_{e_i,0}$ for binary outcome, where e_i denotes the ith canonical vector, i.e. the vector with a 1 in the i-th coordinate and 0's elsewhere.

In both examples, we generate the data with a 1-dimensional latent confounder ($m = 1$), $k = 4$ treatments, $B = [2, 0.5, -0.4, 0.2]$, $\sigma^2_{t|u} = 1$, $\gamma = 2.8$, $\sigma^2_{y|t,u} = 1$, $h_T(\tilde{T}) = \tilde{T}$ and

$$g(T) = 3T_1 - T_2 + T_3 I_{T_3>0} + 0.7 T_3 I_{T_3 \leq 0} - 0.06 T_4 - 4T_1^2.$$

Based on the choice of g, contrasts along the jth dimension of T have effects of widely varying magnitude. Based on our choice for B, the worst-case confounding bias also



(a) Gaussian Outcome (b) Binary Outcome

Figure 4.4: Estimated ignorance region for e_i in case when $\tilde{Y}$ is nonlinear in T. (a) $R^2_{\tilde{Y}\sim U|T} = 0$ denotes the treatment effects estimated based on the observed data only, i.e., under the assumption of no confoundedness. $R^2_{\tilde{Y}\sim U|T} = 1$ and $R^2_{\tilde{Y}\sim U|T} = 1$ correspond to the case when all residual variation in Y is due to the confounding, respectively denoting the upper and lower bounds of the ignorance region. (b) In the binary setting, even though the estimand is a non-linear function of the latent Gaussian outcome, the width of the ignorance region and general robustness pattern is largely consistent with the implications of Corollary 4.4.1.

varies significantly across contrasts. For example, the effect of confounding is larger when estimating the treatment of T_1, since the first entry of B has the largest magnitude, meaning T_1 is the feature most correlated with U. In order to demonstrate this in simulation, we first apply probabilistic PCA (PPCA) to estimate the distribution $f(u \mid t)$, and then model $f(y \mid t)$ using Bayesian Additive Regression Tree (BART) with R package BART [29].

For Gaussian outcomes, the width of the ignorance regions are larger for the treatments most correlated with confounders as characterized Theorem 4.4.3 (see Figure 4.4). Since B is a vector, the width of the ignorance region of PATE$_{t_1,t_2}$ can be examined by looking at the dot product between B and the treatment contrasts. The larger the dot product, the wider the ignorance region. As expected, the ignorance region of the treatment effect is widest when $t_1 = e_1$ (RV $\approx 0\%$) and narrowest when $t_1 = e_4$, since $B'e_1$ has the largest magnitude while $B'e_4$ has the smallest. Despite the fact that $t_1 = e_4$

has the smallest ignorance region, it is not robust to confounding because the naive effect is already close to zero (RV = 9%) . For the second and third treatment contrasts, estimates are robust to confounders, as their entire ignorance regions exclude 0.

For the simulation with binary outcomes, we compute ignorance regions for the risk ratio. Although we do not have a theoretical result about the ignorance regions of the risk ratio, the general trends in the size of the ignorance region and the robustness of effects are comparable to the Gaussian. Most notably, the treatments with the largest ignorance regions are still those which are most correlated with the confounder. On the other hand, because the outcome is non-linear in U, the naive estimate is not at the center of the ignorance region (Figure 4.4b). In fact, the ignorance region is also non-monotone in $R^2_{\tilde{Y}\sim U|T}$ because the variance of the intervention distribution also depends on γ. In this case, one of the endpoints of the ignorance region corresponds to $R^2_{\tilde{Y}\sim U|T} = 1$ but the other does not. We compute the endpoints of the ignorance region numerically (see Appendix A.2.3 for more details).

4.6.2 Example with Simulated Genome Wide Association Study

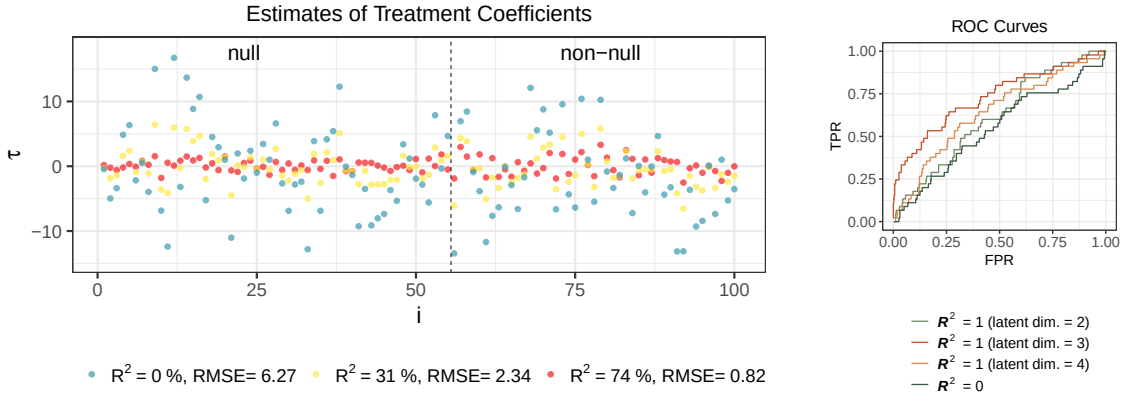
We now explore a slightly more complex setting motivated by applications in biology, particularly in genome wide association studies (GWAS). GWAS investigate the association between hundreds or thousands of genetic features (i.e., single nucleotide polymorphisms, or SNPs) and observable traits (i.e., phenotypes), such as disease status. Here, we construct a simulated GWAS to demonstrate two properties of our sensitivity analysis method. First, we show that flexible latent variable models can be plugged into our sensitivity model. Secondly, we demonstrate how minimizing multiple contrast criteria (MCC) can be used to select interesting candidate models that conform to broad hypotheses about the nature of genetic effects.

Despite having "association" in the name, GWAS is a particularly interesting application area for multi-treatment causal inference. In practice, measures of association in GWAS are often adjusted to afford a causal interpretation in which conclusions speak to how a phenotype would change if the genome were intervened upon. For example, most analyses adjust for "population structure", which correspond to broad genetic patterns induced by population dynamics that are often confounded with geography, ancestry, environment, and other lifestyle factors [42, 63]. Wang and Blei (2018) [36] cite this literature as motivation for their work.

In this simulation, we generate data with high-dimensional binary treatments (SNPs), and set the true causal effects to be mostly small, with a small fraction of treatments having effects of larger magnitudes. The simulation is then designed so that unobserved confounding biases estimates for each of these treatment effects, obscuring the difference between large and small effects. To generate data, we follow the template in Equations 4.41–4.45. We generate data with $m = 3$ latent confounders and $k = 500$ treatments, $T \in \{0, 1\}^k$, where $T_j = 1$ if the the jth site shows a deviation from the baseline sequence (i.e., the presence of at least one minor allele). We set the response function $g(T) = \tau' T$ to be linear in the treatments (a common assumption in GWAS), and set the outcome Y to be Gaussian by setting $h_Y(\tilde{Y}) = \tilde{Y}$. We focus on estimating

$$\frac{1}{n} \sum_{i=1}^{n} PATE_{t_i^j, t_i^{-j}} \text{ for all } j = 1, \cdots, k, \tag{4.46}$$

where t_i^j and t_i^{-j} correspond to the i^{th} observed treatment vector with the j^{th} SNP set to be 1 and 0 respectively. Note that since $g(T)$ is linear in T, $\frac{1}{n} \sum_{i=1}^{n} PATE_{t_i^j, t_i^{-j}} = \tau_j$, the j^{th} element of τ. We generate τ from a two component mixture with 90% of the coefficients from a Uniform$(-0.1,\ 0.1)$ (small effects) and 10% from a Uniform$(-2,\ 2)$ (large effects). We assume that there are $m = 3$ latent confounders.



(a) L1 minimizing effects (b) Non-null classification

Figure 4.5: Causal inference with 500 binary treatments with $k = 3$ latent confounders. Naive estimates of the null and non-null effects are overdispersed due to confounding. (a) Minimum L1-norm treatment effects are shown for fifty randomly chosen small effects ("null" contrasts) and all large effects ("non-null" contrasts) for three different limits on the magnitude of confounding, $\mathcal{R}^2 \in \{0, 0.3, 1.0\}$. When $\mathcal{R}^2 = 1$, there overall L1 minimizer of the treatment effects is achieved for the sensitivity vector which explains $R^2_{Y\sim U|T} = 74\%$ of the residual outcome variance. (b) We construct a simple non-null classifier from minimum L1 treatment effects with $\mathcal{R}^2 = 1$ and naive effects ($\mathcal{R}^2 = 0$). The blue curve represents the ROC curves from the naive estimates and the green, yellow and red curves represents the L1 minimizer of the treatment effect estimates for inferred confounder models with dimensions $\hat{k} \in \{2, 3, 4\}$. The area under the curve (AUC) for the naive estimates is 0.54, whereas the AUC for the L1-minimized estimates are 0.61 ($\hat{k} = 2$), 0.73 ($\hat{k} = 3$) and 0.64 ($\hat{k} = 4$).

We consider a model for the observed data with two components, paying special attention to the latent confounder model. In particular, we model the conditional distribution of confounders given treatment $f(u \mid t)$ using a variational autoencoder (VAE), which is a popular, flexible neural network–based approximate latent variable model. This model is particularly appropriate because it yields an approximate Gaussian conditional distribution $f(u \mid t)$, even for discrete T as we have here. (We discuss latent confounder inference with VAEs in more detail in Appendix A.2.2.) We fit the observed outcome model $f(y \mid t)$ using a simple linear regression, ignoring confounding, which corresponds to the setting in which $R^2_{Y \sim U|T} = 0$.

Worst-Case Ignorance Regions. With this simulation setup, we first examine whether the ignorance regions contain the true causal effects. Importantly, because the VAE is an approximate latent variable model, and we are currently ignoring estimation uncertainty, it is not immediate that the ignorance regions should be valid. We find that, even using our plug-in approach, the worst case ignorance regions cover 498 out of 500 of the true treatment effects. In all cases, the worst case bounds communicate substantial fundamental uncertainty about the true treatment effects (See Appendix Figure A.2).

Finding Candidate Models with MCCs. Investigators often have strong hypotheses about the aggregate properties of SNP treatment effects. For example, while some phenotypes can be predominantly explained by only a small number of SNPs, other phenotypes may be more plausibly described by the omnigenic hypothesis, which suggests that some observable effects must be explained by the sum of many small effects across many SNPs [64]. Here, we show that some of these aggregate hypotheses can be formalized in terms of MCCs, and in these cases, the MCC minimization procedure from Section 4.5.4 can be used to find useful candidate causal models that align with these

hypothesis while being fully consistent with the observed data.

To motivate candidate model selection, we consider the use case of estimating effect sizes from a single coherent model, under the hypothesis that the median effect size is small. Specifically, we formalize this hypothesis by defining a MCC $\omega(\mathbf{PATE}_{\mathcal{T}^2}(\gamma))$ to be the L_1 norm of the effects of each contrast $\mathcal{T}^2 = \{(t_i^j, t_i^{-j}) : i \in (1, ..., n)\}$ for all treatments $j = 1, \cdots, k$. We then select the model that minimizes this criterion by selecting γ subject to different allowed levels of confounding $R^2_{Y\sim U|T}$.

In Figure 4.5a, we plot the the resulting coefficients estimates for three values of $\mathcal{R}^2$: 0 (naive effects), 0.3 and 1. Because the true effects are much smaller in magnitude than the naïve effects, the RMSE of the estimates decreases as we increase $R^2_{Y\sim U|T}$, although all effects are equally compatible with the observed data. In this simulation, the L1 norm of naive estimates is approximately 2525 and the norm of the true effects is drastically smaller at approximately 75.

Models selected using this MCC minimization procedure are also useful for the coarser goal of separating small and large effects. From the naive regression, the coefficients are overdispersed to the true causal effects and the true small coefficients are practically indistinguishable from true large coefficients. Meanwhile, models chosen with the MCC minimization procedure provide more useful signal. To formalize this, we consider a classifier that separates large and small effects using the magnitude of the inferred coefficients as the classification score. In Figure 4.5b we plot the receiver operating characteristic (ROC) curves for the classifiers based on the naive estimates as well as the overall $L1$ minimizer of the treatment effects ($\mathcal{R}^2 = 1$, i.e. no limit on the value $R^2_{Y\sim U|T}$).

Importantly, the difference in conditional confounder means, $\mu_{u|t_i^j} - \mu_{u|t_i^{-j}}$, varies between non-null and null contrasts. This leads to a larger reduction in the relative magnitude of the null effects for models chosen through MCC minimization, accentuating the differences between large and small treatment effects (See Appendix Figure A.3). For

models selected by MCC minimization, the area under the ROC curve (AUC) increases from 0.54 (almost no ability to distinguish small and large treatments) to 0.72 ($\hat{k} = 3$, red curve). The selected model achieves nearly 25% true positive rate without accruing any false positives. Naturally, the classifier performance is the best when we fit a latent variable model with the correct number of latent factors, although the classifier based on latent variable models of dimensions $\hat{k} = 2$ and $\hat{k} = 4$ still outperform classification from naive effects. In the Discussion, we note how this approach relates to, and complements recent identification results for a similar setting in Miao et al. (2020) [45].

4.7 A Reanalysis of the Actor Case Study

In this section, we compare our approach to other recent analyses of the TMDB 5000 Movie Dataset [65] which was analyzed extensively by Wang and Blei (2018) [36] and Grimmer et al. (2020) [41]. The dataset consists of 5000 movies and their corresponding revenue, budget, genre and the identities of the lead cast members. Following Wang and Blei, we focus on estimating the causal effect of an actor's presence on the movie's log revenue. We let Y denote the log revenue and $T_i = (T_{i1}, \cdots, T_{ik})$ encode the movie cast, where the binary random variable $T_{ij} \in \{0, 1\}$ indicates whether actor j appeared in the movie i and $T_i \in \mathcal{T} = \{T_1, ..., T_n\}$. We also let $\mathcal{T}^j$ denote the set of all movies T_i for which $T_{ij} = 1$. We define the estimand of interest, η_j, as the the total log revenue contributed by actor j:

$$\eta_j := \sum_{t_i \in \mathcal{T}^j} \text{PATE}_{t_i, t_i^j} \tag{4.47}$$

where t_i^j corresponds to the observed treatment vector for movie i excluding actor j. This estimand is a non-parametric generalization of the regression coefficient τ_j, which was

targeted in the analysis in Wang and Blei (2018) [36]. Specifically, under the assumption that log-revenue is linear in the cast indicators, η_j reduces to $n_j\tau_j$, the effect of actor j scaled by the number of movies they appeared in, where τ_j are the regression coefficients for actor j. Our estimand is well-defined without this linearity assumption.

We regress the log revenue on cast indicators to estimate actor effects, τ_j^{naive}, under an assumption of no unobserved confounding. In order to demonstrate the applicability of our sensitivity analysis, we explicitly induce unobserved confounding by excluding observed confounders. We validate our analysis, by comparing calibrated effect estimates when the confounders are excluded to estimates when the confounder is included. Most importantly we exclude the movie's budget which we estimate to be the largest known source of confounding (computed using Equation 4.32, see Appendix Figure A.4a)[3].

For simplicity, we model the observed outcome distribution with a linear regression, although other more flexible outcome models (e.g. `BART`) can also be used. As in the previous Section, we use a VAE to infer a Gaussian conditional confounder distribution, $f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t})$ (See Appendix, Section A.2.2).

Results. Since our focus is on confounding not estimation, in order to limit the influence of estimation uncertainty we subset the data to the $k = 327$ actors who participated in at least twenty movies. This reduces the total number of movies to 2439. We fit the VAE to the treatments and use cross-validation to identify the appropriate latent confounder dimension, which we inferred to be $\hat{m} = 20$ (See Appendix Figure A.4b). We then plot the worst-case ignorance region for the causal effect on log revenue as a function of $R^2_{Y\sim U|T}$ for the 46 actors with significant regression coefficients in the naive regression (Figure 4.6, top). Eight actors in the observed data regression have significantly negative coefficients, whereas 38 actors have significant positive coefficients. However, the worst-

[3]For illustrative purposes, we can assume that the budget is pre-treatment, meaning that the budget is decided prior to selecting the cast, which may be a dubious assumption in actuality.

case ignorance regions for each actor are all very wide and include zero, which suggests that none of the effects are robust to confounding. In Table A.1 of the Appendix we include robustness values for these actors. Leonardo DiCaprio has the largest robustness value at 36%, with the majority of the other actors well below 20%. For reference, the log budget, which was explicitly excluded from our causal analysis, explains about 30% of the variance in log revenue (Appendix Figure A.4a). In other words, none of the causal effects are robust at a level which matches the variance explained by the most important excluded confounder.

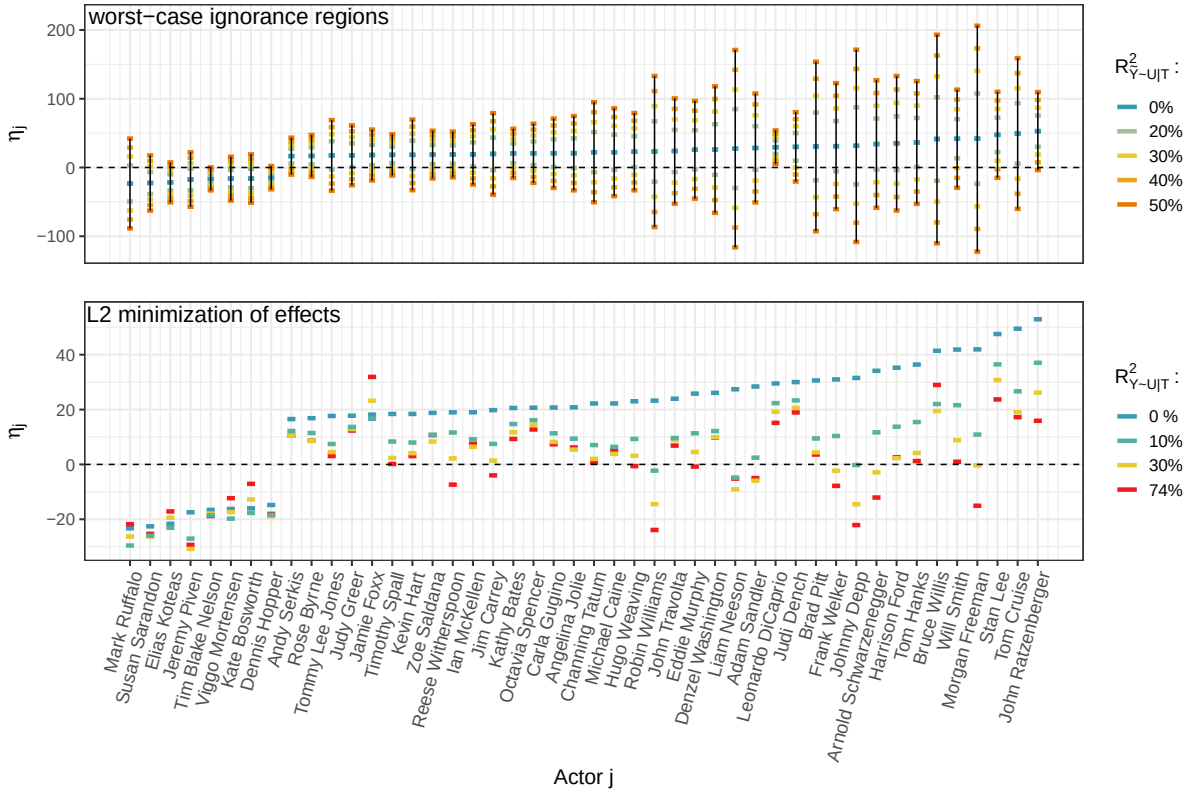


Figure 4.6: Estimated total log revenue contributed by a given actor. Top: worst-case ignorance region for each actor on a case by case basis. The blue points correspond to $R^2_{Y\sim U|T} = 0$, i.e. the naive estimates. Robustness values can be found in Appendix Table A.1. Bottom: Treatment effects for candidate models chosen with the L2 minimzing multiple contrast criterion (MCC). The color correspond to $\mathcal{R}^2$, the limit on the fraction of outcome variance explained by confounding.

The worst case ignorance regions depicted in top panel of Figure 4.6 correspond to a

different choice of γ for each actor. We can also explore the robustness of causal effects under a single model by applying an appropriate MCC. Specifically, we search for a "worst case" candidate model by finding the sensitivity vector, γ_*, that implies the smallest L2 norm of the regression coefficients, τ. In this conservative model, the minimum L2 norm of the treatment coefficients is 4.4, down from 7.6 for the naive coefficients. In addition, 40 out of 46 actors have coefficients that are smaller in magnitude than the magnitudes of the naive coefficients (Figure 4.6, bottom). For this candidate model, it turns out that $\gamma_*' E[U|T = t]$ is significantly correlated, albeit weakly, with budget (Spearman's rank correlation = 0.2, p-value $< 2e$-16). Thus, the conservative model correctly attributes part of the outcome variation induced by the known excluded confounder to unobserved confounding.

4.8 Discussion

In this chapter, we introduced a framework for sensitivity analysis with multiple treatments which provides further context to the growing literature on the challenges of inference in this setting. Unlike previous work, we emphasize the importance of carefully defined estimands and show that bounds on the magnitude of confounding bias depend on the particular estimands of interest. Our work also provides a practical solution to characterizing and calibrating the robustness of causal effects across multiple treatments in the presence of unobserved confounding. Code to replicate all analyses is available [66] and an R package implementing our methodology is also available and in active development [67].

There are several interesting generalizations of our proposed approach, many of which center on assumptions about the copula. For example, in many contexts we may be interested in accounting for potential treatment-confounder interactions, in which case

the copula, $c(F_{Y|t}(y), F_{U|t}(u) \mid T = t)$, will vary with t. Likewise, as noted, our approach can be applied with non-Gaussian copulas and and alternative latent variable models (e.g latent class models), but calibration and model-specification remains a challenge.

Generalizations based on joint inference of the treatment and outcome models should also be explored. In practice, causal effect estimates may be overdispersed about the true effects due to a combination of sampling variance and unobserved confounding. Joint inference might be especially useful for accounting for both estimation uncertainty and uncertainty due to unobserved confounding. This is particularly important for the multiple contrast criteria which, as described, does not incorporate parameter uncertainty into the objective function. A simple solution in a Bayesian analysis would be to characterize posterior uncertainty by applying the criteria to each MCMC sample of the naive causal effect estimates. A more thorough exploration of the interplay between shrinkage estimation (in the classical sense) and MCC shrinkage to adjust for confounding bias under the "small effects" hypothesis would be interesting to explore in this context.

Finally, we note that there are many promising directions for incorporating additional constraints into the calibration criteria, besides the null control treatments. Incorporating these assumptions would further constrain the ignorance regions for particular causal contrasts of interest. These strategies are closely related to the multivariate calibration procedure that we propose. For example, Miao et al. (2020) [45] describe a procedure for identifying the treatment effects when over half of the treatments are assumed to have no causal effect on the outcome. In Equation 4.40, this would be analogous to the case in which m is the L_0 norm, the number of non-zero effects. It is also worth further exploring the relationship between inference with multiple treatments and inference with multiple outcomes. Additional structure in the correlation between outcomes could further constrain the causal ignorance regions, under the right set of assumptions. As with null control outcomes, which are known to have null treatment effects, could be applied

to calibrate the sensitivity analysis. We leave this to future work.

Chapter 5

Copula-Based Sensitivity Analysis with Multiple Outcomes

Many practical causal inference problems also involve high-dimensional outcomes, for example, observational studies examining the effect of a treatment on multiple biomarkers [68, 69, 70]. Similar to the multi-treatment case, the multivariate correlation structures can also provide additional implications about the unobserved confounders. Building on previous results, we develop a copula-based sensitivity analysis for cases with multiple correlated outcomes in this chapter. Unlike the multi-treatment setting, we show that we cannot bound the treatment effects, although in practice, the ignorance regions of treatment effects don't blow up unless there is extremely large confounding. We also propose calibration strategies including calibrating the sensitivity parameters, quantifying robustness of effect estimates, and, in particular, incorporating prior knowledge about outcomes that have null treatment effects to further constrain the ignorance region.

5.1 Introduction

In classical causal inference studies, the goal is often to find out the impact of a treatment or intervention on a single outcome. However, causal inference on multiple non-independent outcomes is increasingly widespread in real-world applications. Especially, simultaneous effects of a treatment on multiple outcomes may be of particular interest, such as, in biological studies, the effects of high fish assumption on multiple biomarkers related to organochlorines [70], or in patient-centered epidemiologic studies, the prioritization of public health recommendations [71, 72]. Multi-outcomes cases may also arise when we are interested in the causal effects from multiple aggregate time series before and after an intervention, for example, when there are state-level policy changes or the introduction of a new marketing campaign [73].

A strand of causal inference literature has considered causal identification in multi-outcome settings with intermediate variables, but still focusing on causal effects on a single response and using other outcome variables as auxiliary variables for inference [74, 75, 76, 77]. VanderWeele and others [71, 72, 78] suggest that, instead of using one-outcome-at-a-time approach, more effort should be made on developing multivariate techniques that account for the dependence structure of outcomes, and exploring effects of interventions on multiple outcomes simultaneously. In more recent works, researchers have begun focusing on assessing causal effects on multivariate response variables. Lupparelli, M., & Mattei, A. (2017) [78] consider cases of binary outcomes, and decompose the treatment effects of multivariate outcomes into the joint and marginal causal effects, which respectively provide information on the marginal and dependent structure of the outcomes, and propose a log-mean linear regression for modeling the outcome distribution that can account for the outcome's correlation structure. Kennedy, E. H., Kangovi, S., & Mitra, N. (2019) [71] develop a doubly robust method for estimation and hypoth-

esis testing of scaled treatment effects on multiple outcomes. More studies for causal inference problems with multivariate outcomes can be found in [79, 80, 81, 82]. A major difficulty of these aforementioned works is that they are derived under a version of the "no unobserved confounding" assumption. As mentioned, this is often a dubious assumption in practice. For example, policy interventions can be motivated by anticipated future effects: a company may be more likely to adopt a marketing campaign if expected future sales were to be particularly bad in the absence of such a campaign. Therefore, it would also be of great value to develop sensitivity analysis methods for multi-outcome models, especially considering that, to our best knowledge, there is no sensitivity analysis approach developed particularly for causal inference problems with multivariate outcomes in the literature.

In this chapter, utilizing the copula-based framework developed in Chapter 4, we propose a sensitivity analysis method for causal inference problems with multivariate outcomes As in the multi-treatment case, the copula-based factorization can be used to clearly separate the unidentifiable part from the identifiable part. As a special case of the general model, we demonstrate our sensitivity analysis method with Gaussian copulas, which is able to capture essential qualitative aspects of confounding in the multi-outcome cases. As we will show, unlike the multi-treatment settings, we cannot globally bound our ignorance about the treatment effect with multiple outcomes. However, the multivariate correlation structure of the outcomes still provides important additional information about confounders that, in practice, the ignorance region of the treatment effects don't blow up unless there are extremely large confounders. In addition to the calibration methods discussed in 4.5, we provide strategies that incorporate prior knowledge about outcomes with null effects, for further reducing the ignorance regions.

The rest of this chapter is organized as follows. In Section 5.2, we set up the basic framework for copula-based sensitivity analysis with multivariate outcomes. In Section

5.3, we provide some theoretical insight into confounding bias under a special case of our general model, where we assume the relationship between the confounders and the treatment as well as the outcomes can be characterized by Gaussian distributions. In Section 5.4, we discuss the interpretation and calibration of sensitivity parameters, including strategies that can account for additional prior information about outcomes with null effects. Finally, in Section 5.5, we demonstrate our approach with an analysis of metabolomic aging clocks studied in [83].

5.2 Sensitivity Analysis via Copula Parameterizations with Multi-Outcome

In the multiple outcome settings, unlike the multi-treatment cases, the conditional confounder distribution $f_{\psi_T}(u \mid t)$ is not identifiable, but, instead, the conditional outcome distribution $f_{\psi_Y}(y \mid t, u)$ can be partially identified from the multivariate correlation structure of the outcomes. We focus on this specific case where the relationship between confounders and outcomes are partially identifiable under appropriate assumptions.

In this section, we discuss how copula-based factorization can be applied to the multiple outcome setting, and elaborate its implementation with the Gaussian copula in particular.

5.2.1 Multiple Outcomes and Causal Equivalence Classes

In the multi-outcome settings, besides the causal estimands defined in Section 1.2, which are all vectors, another commonly used estimand is

$$\text{PATE}_{a,t_1,t_2} := E(a'Y \mid do(t1)) - E(a'Y \mid do(t2)), \tag{5.1}$$

which is a scalar and stands for the treatment effect of treatment on a linear combination of the outcome, $a'Y$.

As discussed in Section 4.2.1, the intervention distribution $f(y \mid do(t))$ can be generally factorized in terms of copula as

$$f_{\psi}(y \mid do(t)) = f(y \mid t) \int c_{\psi_Y}(F_{Y|t}(y), F_{U|t}^{\psi_T}(u) \mid t) f(u) du,$$

There are no observable implications on sensitivity parameters, and the relationships of outcome-confounder and treatment-confounder are parameterized separately. With multiple outcomes and single treatment, the conditional confounder distribution $f_{\psi_T}(u \mid t)$ is unidentifiable, however, the relationship between outcomes and confounders conditional on treatment, parameterized by ψ_Y, is often identifiable up to causal equivalence class.

Definition 5.2.1 (Causal equivalence class in multi-outcome setting) *$[\psi_Y]$ is a causal equivalence class of ψ_Y if and only if for any $\tilde{\psi}_Y$ in $[\psi_Y]$, then, for every ψ_T there exists a $\tilde{\psi}_T$ such that $f_{\psi_Y,\psi_T}(y \mid do(T=t)) = f_{\tilde{\psi}_Y,\tilde{\psi}_T}(y \mid do(T=t))$ for all y, t.*

This definition implies that when we say ψ_Y is identifiable up to a causal equivalence class, it means that the causal conclusions don't change with value of ψ_Y within the class. In the following of this chapter, we focus on scenarios where ψ_Y can be identified up to a causal equivalence class. To conduct sensitivity analysis, we only need to assume that ψ_Y is point-identified at a particular value within the class $[\psi_Y]$.

5.3 Sensitivity Analysis with Multiple Outcomes in the Gaussian Copula Model

We again assume c_ψ is a Gaussian copula that it is invariant to the level of T (Assumption 4.3.2 and 4.3.1). As mentioned in Section 4.3, Gaussian copula facilitates interpretation and sensitivity parameter calibration without imposing any restrictions on the observed data distributions, $f(y \mid t)$ and $f(t)$. Assumption 4.3.1 and 4.3.2 together imply that the conditional confounder follows a Gaussian distribution with covariance matrix invariant to the level of treatment. In the multi-outcome model, this implies the following generative model:

$$T \sim F_T,\ \mathrm{E}(T) = \mu_t,\ \mathrm{Var}(T) = \sigma_t^2 \tag{5.2}$$

$$f(u \mid t) \sim N\Big(\frac{\Sigma_u \beta}{\sigma_t^2}(t - \mu_t),\ \Sigma_u - \frac{\Sigma_u \beta \beta' \Sigma_u}{\sigma_t^2}\Big) \tag{5.3}$$

$$\tilde{Y} = [\tilde{Y}_1, \cdots, \tilde{Y}_q]' = \Gamma(U - \mu_{u|t}) + \epsilon_{\tilde{y}|t,u}, \tag{5.4}$$

$$\epsilon_{\tilde{y}|t,u} \sim N_q(0, \Lambda_{\tilde{y}|t,u}),\ \Gamma \Sigma_{u|t} \Gamma' + \Lambda_{\tilde{y}|t,u} = C_{y|t}, \tag{5.5}$$

$$Y = [Y_1, \cdots, Y_q]' = [F^{-1}_{Y_1|t}(\Phi(\tilde{Y}_1)), \cdots, F^{-1}_{Y_q|t}(\Phi(\tilde{Y}_q))]', \tag{5.6}$$

where F_T is the CDF of the treatment, $F^{-1}_{Y_i|t}$ is the inverse-CDF of the conditional distribution of Y_i given $T = t$ for $i = 1, \ldots, q$, $\Lambda_{\tilde{y}|t,u}$ is an arbitrary diagonal matrix, and $C_{y|t}$ denotes the correlation matrix of the observed outcome distribution. The Gaussian copula, $c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u) \mid t)$, can be parameterized by the correlation matrix implied by

$$Cov([\tilde{Y}, U] \mid t) = \begin{bmatrix} C_{y|t} & \Gamma \Sigma_{u|t} \\ \Sigma_{u|t} \Gamma' & \Sigma_{u|t} \end{bmatrix}, \tag{5.7}$$

where $\Sigma_{u|t} = \Sigma_u - \frac{\Sigma_u \beta \beta' \Sigma_u}{\sigma_t^2}$ as defined in Equation 5.3, and parameters are $\psi_T = \{\beta\}$ and $\psi_Y = \{\Gamma\}$. In general, Γ is not point identifiable, but, under many latent confounder models, it can be identified up to an invertible linear transformation of U. The following theorem states that the class of ψ_Y defined by all invertible linear transformation of U is a causal equivalence class.

Theorem 5.3.1 *Assume model 5.9-5.6. Let* $[\psi_Y] = \{\tilde{\psi}_Y = \{\Gamma A\} : A \in \mathcal{S}^+\}$ *where* $\mathcal{S}^+$ *is the space of symmetric positive definite matrices. Then* $[\psi_Y]$ *is a causal equivalence class.*

Proof. See appendix.

Theorem 5.3.1 guarantees that for any invertible linear transformation of confounder, $A^{-1}U$, the sensitivity parameter $\tilde{\beta} = A^T\beta$ leads to the same causal conclusions in the reparameterized coordinates of confounder as β does in coordinates of the original confounder U. Throughout this chapter, we will assume that ψ_Y is identified up to invertible linear transformations of U, and explore the range of possible causal effects for different β satisfying $\beta^T\Sigma_u\beta \leq \sigma_t^2$.

5.3.1 Prototype: Linear-Gaussian Model

We begin by illustrating our method in a simple Linear-Gaussian model where (Y, T, U) are jointly multivariate Gaussian. Specifically, the model is specified as following:

$$U = \epsilon_u, \quad \epsilon_u \sim N_m(0, \Sigma_u), \tag{5.8}$$

$$T = \beta' U + \epsilon_{t|u}, \quad \epsilon_{t|u} \sim N_k(0, \sigma_{t|u}^2), \tag{5.9}$$

$$Y = \tau T + \Gamma U + \epsilon_{y|t,u}, \quad \epsilon_{y|t,u} \sim N(0, \Lambda_{y|t,u}), \tag{5.10}$$

with $\beta \in \mathbb{R}^m$, $\tau \in \mathbb{R}^q$ and $\Gamma \in \mathbb{R}^{q \times m}$. Note that model 5.8-5.10 is a special case of model 5.2-5.6, where the association between outcomes and confounders conditional on the treatment, parameterized by Γ, can be characterized by Gaussian copula. Under model 5.8-5.10, the intervention distribution density has

$$f(y \mid do(T = t)) \sim N_q(\tau t,\ \Lambda_{y|t,u} + \Gamma \Sigma_u \Gamma'). \tag{5.11}$$

The observed outcome distribution can be expressed as

$$f(y \mid T = t) \sim N_q(\tau^{\text{naive}} t,\ \Sigma_{y|t}), \tag{5.12}$$

where

$$\sigma_t^2 = \beta' \Sigma_u \beta + \sigma_{t|u}^2, \tag{5.13}$$

$$\tau^{\text{naive}} = \tau + \frac{\Gamma \Sigma_u \beta}{\sigma_t^2}, \tag{5.14}$$

$$\Sigma_{y|t} = \Lambda_{y|t,u} + \Gamma(\Sigma_u - \frac{\Sigma_u \beta \beta' \Sigma_u}{\sigma_t^2})\Gamma', \tag{5.15}$$

$$= \Lambda_{y|t,u} + \Gamma \Sigma_{u|t} \Gamma', \tag{5.16}$$

which are all fully identifiable based on the observed data. As before, τ^{naive} refers to the estimate of τ that naively neglects potential unobserved confounders. Equation 5.13 shows that the marginal treatment variance can be decomposed into non-confounding variance $\sigma_{t|u}^2$ and confounding variance $\beta' \Sigma_u \beta$, where the confounding variance is constrained by the overall magnitude of the marginal variance of the treatment (identifiable). We view σ_t^2, τ^{naive}, $\Sigma_{y|t}$ and therefore $\text{Cor}(Y \mid t) := C_{y|t}$ as fixed and known, and explore how the confounding changes as a function of B.

In the linear Gaussian model, PATE_{a,t_1,t_2} is linear in the difference between two

treatments, $t_1 - t_2$. Therefore, we can assume that $t_1 - t_2 = 1$ without loss of generality and PATE_{a,t_1,t_2} equals

$$\text{PATE}_a := a'\tau(t_1 - t_2) = a'\tau, \tag{5.17}$$

which is invariant to the exact level of t_1 and t_2. The confounding bias of τ^{naive}, $\text{Bias}_a = a'\tau^{\text{naive}} - \text{PATE}_a$, can then be expressed as

$$\text{Bias}_a = \frac{a'\Gamma\Sigma_u\beta}{\sigma_t^2}. \tag{5.18}$$

As show in Chapter 4 previously, in the univariate setting where both the treatment and outcome are single, neither ψ_T and ψ_Y are identifiable, while, in the multiple treatment setting, ψ_T is identifiable up to a causal equivalence class. Analogously, in the multiple outcome setting, ψ_Y is identifiable up to a causal equivalence class defined by invertible linear transformation of U in the multiple outcome setting. We formalize the idea in the following theorem.

Theorem 5.3.2 *Suppose that the observed data is generated by model 5.8-5.10. When there are q outcomes with $1 < m < q$, then ψ_Y is identified up to the causal equivalence class $[\psi_Y] = \{\tilde{\psi}_Y = \{\Gamma A\} : A \in \mathcal{S}^+ \}$. When there is a single outcome ($q = 1$) or at least $m = q$ confounders, then ψ_Y is not identifiable up to causal equivalence class.*
Proof. See appendix.

Theorem 5.3.1 and 5.3.2 indicate that the distribution of U is only causally relevant up to linear transforms, and as such, without loss of generality, we make the simplifying assumption that $U \sim N(0,\ I_m)$ throughout the remainder of this chapter. Plugging in

$\Sigma_u = I_m$, the confounding bias equals

$$\text{Bias}_a = \frac{a'\Gamma\beta}{\sigma_t^2} \tag{5.19}$$

$$= \frac{1}{\sigma_t^2} a'\tilde{\Gamma}(I_m - \frac{\beta\beta'}{\sigma_t^2})^{-1/2}\beta, \tag{5.20}$$

where $\tilde{\Gamma} := \Gamma\Sigma_{u|t}^{1/2}$ is identifiable up to a causal equivalence class. The relationship between treatment and confounder, parameterized by $m-$vector β, remains unidentified. Although β is not identified, its magnitude is constrained by the identifiable treatment variance. It must satisfy the constraint,

$$0 \leq R^2_{T\sim U} = \frac{\beta'\beta}{\sigma_t^2} < 1, \tag{5.21}$$

which is the fraction of variation in the treatment due to confounding. Importantly, the upper bound of $R^2_{T\sim U}$ indicates that the confounding variation in T needs to be *strictly* less than the total variation of treatment, which ensures the positivity condition (Assumption 1.3.2). The constraint on β implies the following result about the confounding bias of the naive estimate.

Theorem 5.3.3 *Suppose that the observed data is generated by model 5.8-5.10. Then,* $\forall\beta$ *satisfying*

$$\beta'\beta = \sigma_t^2 R^2_{T\sim U} \tag{5.22}$$

with $0 \leq R^2_{T\sim U} < 1$. *For a given vector* a, *the confounding bias is bounded by*

$$Bias_a^2 \leq \frac{1}{\sigma_t^2}\frac{R^2_{T\sim U}}{1 - R^2_{T\sim U}} \parallel a'\tilde{\Gamma} \parallel_2^2, \tag{5.23}$$

where the bound is achieved when β *is collinear with* $a'\tilde{\Gamma}$.

Proof. See appendix.

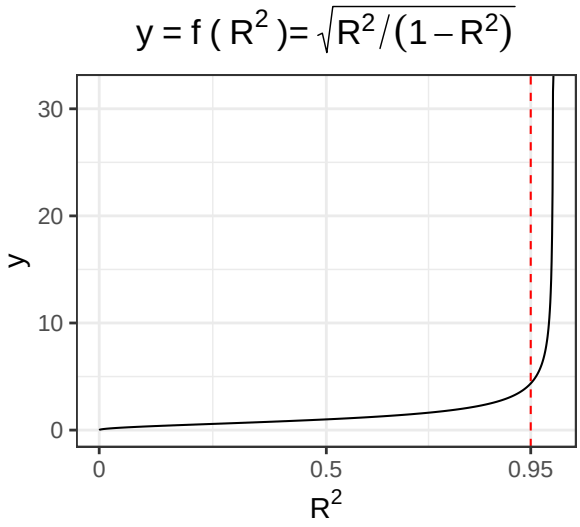


Figure 5.1: Plot the factor of confounding bias that only depends on $R^2_{T\sim U}$. The value of $\sqrt{R^2/(1-R^2)}$ doesn't blow up until R^2 is larger than about 0.95.

This theorem states that the true treatment effect lies in the interval $a'\tau^{\text{naive}} \pm \sqrt{\frac{1}{\sigma_t^2}\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}}$ $\| a'\tilde{\Gamma} \|_2$. As in the previous chapter, we refer to the right-hand side of Equation 5.23 as the "r^2%-$R^2_{T\sim U}$ bias" of the naive estimator given the value of $R^2_{T\sim U}$ as r^2%. From Equation 5.23, the r^2%-$R^2_{T\sim U}$ bias depends on $\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}$, the ratio between the confounding variation and non-confounding variation in T. $\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}$ scales the bias for all estimands by an equal proportion. Importantly, in contrast to the multi-treatment setting, the bias about the treatment effect is unbounded, since $\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}$ can be arbitrary large. But, in practice, the ignorance regions of treatment effects don't explode unless $R^2_{T\sim U} > 0.95$ (see Figure 5.1). The last factor of the confounding bias, $\| a'\tilde{\Gamma} \|_2^2$, measures the degree to which $a'Y$ depends on the confounder U and varies across outcomes of interest. In the following corollary, we characterize which outcomes have the largest bias and which outcomes have identifiable treatment effects under Assumptions 1.3.1, 1.3.2, 4.3.2 and 4.3.1.

Corollary 5.3.1 *Let d_1 be the largest singular value of $\tilde{\Gamma}$. For all $a \in \mathcal{R}^q$ with $\| a \|_2 = 1$,*

the confounding bias is bound by

$$Bias_a^2 \leq \frac{d_1^2}{\sigma_t^2} \frac{R^2_{T\sim U}}{1 - R^2_{T\sim U}}, \tag{5.24}$$

with equality when $a = u_1^{\tilde{\Gamma}}$*, the first left singular vector of* $\tilde{\Gamma}$*, and* β *being collinear with* $v_1^{\tilde{\Gamma}}$*, the first right singular vector of* $\tilde{\Gamma}$*. When* $a \in Null(\tilde{\Gamma})$*, the naive estimate is unbiased, that is,* $a'\tau^{naive} = a'\tau$*.*

Proof. See appendix.

Here, d_1^2 denotes the variance of the first principal component of the residual outcomes explained by confounding. For $a = u_1^{\tilde{\Gamma}}$, $a'Y$ corresponds to the projection of outcomes that is most associated with confounders, and therefore is the outcome of interest that has the largest ignorance region. For any fixed $R^2_{T\sim U}$, when a is in the null space of $\tilde{\Gamma}$, PATE_a is identified because $a'Y$ is uncorrelated with the confounders.

5.3.2 Generalizing Linear-Gaussian Model

Next, we extend our approach beyond the linear-Gaussian model. First, when Y is Gaussian with arbitrary conditional mean $\mu_{y|t}$ and diagonal conditional variance matrix $D_{y|t}$, we have

$$E(Y \mid do(t)) = \mu_{y|t} - \frac{1}{\sigma_t^2} D_{y|t}^{1/2} \Gamma \beta (t - \mu_t) \tag{5.25}$$

$$= \mu_{y|t} - \frac{1}{\sigma_t^2} D_{y|t}^{1/2} \tilde{\Gamma} (I_m - \frac{\beta\beta'}{\sigma_t^2})^{-1/2} \beta (t - \mu_t), \tag{5.26}$$

where $\tilde{\Gamma}$ is the $p \times q$ matrix satisfying $\tilde{\Gamma}\tilde{\Gamma}' + \Lambda_{\tilde{y}|t,u} = C_{y|t}$, and therefore, we have

$$\text{PATE}_{a,t_1,t_2} = a'(\mu_{y|t_1} - \mu_{y|t_2}) - \frac{1}{\sigma_t^2} a' D_{y|t}^{1/2} \tilde{\Gamma} (I_m - \frac{\beta\beta'}{\sigma_t^2})^{-1/2} \beta (t_1 - t_2), \tag{5.27}$$

with its confounding bias denoted as $\text{Bias}_{a,t_1,t_2} = a'(\mu_{y|t_1} - \mu_{y|t_2}) - \text{PATE}_{a,t_1,t_2}$.

Then, we have the following extension of Theorem 5.3.3

Theorem 5.3.4 *Assume the model 5.2-5.6 with Gaussian outcomes. The confounding bias of* $PATE_{a,t_1,t_2}$ *is bounded by*

$$Bias^2_{a,t_1,t_2} \leq \frac{1}{\sigma_t^2} \frac{\|\beta\|_2^2}{\sigma_t^2 - \|\beta\|_2^2} \parallel a' D_{y|t}^{1/2} \tilde{\Gamma} \parallel_2^2 (t_1 - t_2)^2, \tag{5.28}$$

where the bound is attained when β *is collinear with* $a' D_{y|t}^{1/2} \tilde{\Gamma}$.

$\frac{\|\beta\|_2^2}{\sigma_t^2 - \|\beta\|_2^2}$ can be viewed analogously as $\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}$, the ratio between confounding variation and non-confounding variation in the treatment T. Since $\|\beta\|_2^2$ measures the confounding strength, σ_t^2 is the marginal variance of treatment, and thus the denominator, $\sigma_t^2 - \|\beta\|_2^2$, measures the non-confounding fraction. In Section 5.4, we describe the strategy for calibrating the magnitude of β by leveraging the idea of (implicit) partial R^2 of T explained by U. As before, $\parallel a' D_{y|t}^{1/2} \tilde{\Gamma} \parallel_2^2$, represent the strength of association between U and $a'Y$ conditional on T. When $a'Y$ lies in the direction that has the largest cosine similarity with U, we have the most ignorance about its treatment effect. When $a'Y$ lies in the null space of $\tilde{\Gamma}$, then the outcome of interest is uncorrelated with confounders, and thus the naive estimator is unbiased.

When the observed outcome distribution is non-Gaussian, we cannot necessarily express PATE_{a,t_1,t_2} analytically, but can still calculate it (see Algorithm 2).

5.4 Calibration

In this section, we describe strategies for calibrating β in our sensitivity model. Following the strategies used in the multi-treatment setting, we calibrate our sensitivity

Algorithm 2: Marginal Contrast Estimation with Multiple Outcomes.

Function `ComputeMean`(t, β)**:**
  for $i = 1, 2, \ldots, n$ **do**
    $\mu_i \leftarrow \frac{1}{\sigma_t^2}\tilde{\Gamma}(I_m - \frac{\beta\beta'}{\sigma_t^2})^{-1/2}\beta(t_i - t)$;
    for $j = 1, 2, \ldots, nSim$ **do**
      Sample $\tilde{y}_{ij} = [\tilde{y}_{ij1}, \cdots, \tilde{y}_{ijq}]'$ from $N(\mu_i,\ C_{y|t})$;
      $y_{ij} = [y_{ij}, \cdots, y_{ij}]' \leftarrow [F^{-1}_{Y_1|t}(\Phi(\tilde{y}_{ij1})), \cdots, F^{-1}_{Y_q|t}(\Phi(\tilde{y}_{ijq}))]'$;
  return $\frac{1}{n}\sum_{ij} v(y_{ij})$
return $\tau(\text{ComputeMean}(t_1, \beta), \text{ComputeMean}(t_2, \beta))$

parameter vector β by considering the fraction of treatment variance explained by unobserved confounders. Recall that β is a vector that parameterizes the correlation between m-dimensional unobserved confounder U and treatment T. As argued in Chapter 4, in the univariate case, the causal effect is identified given two sensitivity parameters: the fraction of outcome variance explained by unobserved confounders, $R^2_{Y\sim U|T}$, and the fraction of treatment variance explained by unobserved confounders, $R^2_{T\sim U}$ [18]. In contrast to the multi-treatment setting, there is no need to calibrate $R^2_{Y\sim U|T}$, because we are considering the case where the residual correlation between outcomes and confounders after conditioning on T is identifiable up to a causal equivalence class. Instead of the outcome-confounder relationship, the treatment-confounder relationship, parameterized by β, is unknown. As β is a vector, we separately consider its magnitude and direction.

5.4.1 Calibration in General

Calibrating the magnitude of β. When the treatment is Gaussian, the magnitude of β can be directly characterized by $R^2_{T\sim U}$ since

$$R^2_{T\sim U} = \frac{\text{Var}(T) - \text{Var}(T \mid U)}{\text{Var}(T)} = \frac{\|\beta\|_2^2}{\sigma_t^2} \tag{5.29}$$

is the fraction of variation in T that can be explained by U. Similar to previous chapters, we can calibrate $R^2_{T\sim U}$ by comparing it to the fraction of variance explained by different observed covariates when they are available. Let X denotes observed covariates, and, without loss of generality, assume that X and U are independent. As done in Cinelli and Hazlett (2019) [18], we compute the fraction of variation in Y that can be explained by a specific covariate (or set of covariates), X_j, after controlling for all other covariates X_{-j},

$$R^2_{T\sim X_j|X_{-j}} := \frac{R^2_{T\sim X} - R^2_{T\sim X_{-j}}}{1 - R^2_{T\sim X_{-j}}}. \tag{5.30}$$

For non-Gaussian treatment, calibration is less straightforward. Again, we adopt the idea of "implicit R^2" from Imbens (2003) [3], which is a measurement for non-Gaussian variable, defined in parallel to the ordinary R^2. When the observed treatment is binary, we posit a logistic regression model for the treatment assignment. With Equation 5.2, it can be shown that the coefficient of U equals $\frac{\beta'}{\sigma_t^2 - \|\beta\|_2^2}$ in the logistic treatment model. This implies

$$f(T = 1 \mid U, X) = \text{logit}^{-1}\{m(X) + \frac{\beta'}{\sigma_t^2 - \|\beta\|_2^2}U\}. \tag{5.31}$$

Under model 5.31, the implicit partial R^2 of U in T equals

$$R^2_{\tilde{T}\sim U|X} := \frac{\frac{\|\beta\|_2^2}{(\sigma_t^2 - \|\beta\|_2^2)^2}}{\text{Var}(m(X)) + \frac{\|\beta\|_2^2}{(\sigma_t^2 - \|\beta\|_2^2)^2} + \pi^2/3}, \tag{5.32}$$

which can be compared with the partial R^2 of X_j defined in Equation 5.30 with

$$R^2_{T\sim X} = \frac{\text{Var}(E(T \mid X))}{\text{Var}(E(T \mid X)) + \pi^2/3}. \tag{5.33}$$

For more details, see Imbens (2003) [3] and Franks et al. (2019) [2] who discuss this

strategy of using implicit R-squared values in logistic regression models for calibration.

Calibrating the direction of β. Given a magnitude of $R^2_{\tilde{T}\sim U}$, we now propose a default method for identifying the direction of β for outcome of interest, $a'Y$. By default, we suggest using the direction that maximizes the magnitude of confounding bias. As shown in Theorem 5.3.4, when β is collinear with $a'D^{1/2}_{y|t}\tilde{\Gamma}$, the confounding bias of the naive estimator for Gaussian outcomes is maximized at

$$| \text{ Bias}_a | = \sqrt{\frac{\|\beta\|_2^2}{\sigma_t^2(\sigma_t^2 - \|\beta\|_2^2)}} \parallel a'D^{1/2}_{y|t}\tilde{\Gamma} \parallel_2 . \tag{5.34}$$

As before, for non-Gaussian outcomes or alternative estimands, there may not be an analytic solution to the direction which maximizes the bias, but, with our copula-based method, we can still compute the direction via numerical optimization.

5.4.2 Calibration with Null Control Outcomes

Null control outcomes, which are outcomes that are assummed to be unaffected by the treatment, have long been used to detect and correct potentially unobserved confounding [84, 85, 51]. For example, researchers may wish to use Magnetic Resonance Imaging (fMRI), which measures brain activity by detecting changes associated with blood flow, to learn which regions of the brain are activated by a particular auditory stimulus, where the confoundedness could be induced by the level of blood oxygenation, but voxels in the white matter or cerebrospinal fluid can often be regarded as null controls to adjust for the unwanted confoundedness [85]. Many statistical methods have been developed to identify the true causal effect by using null controls [86, 87, 88, 84, 85]. Instead of focusing on identification under strong assumptions, we show that null control outcomes can be used to further shrink our ignorance about the treatment effect. We illustrate the

idea specifically under model 5.8-5.10 in the following.

Let $\mathcal{C}$ be a set indexing c null control outcomes such that $\tau_j = 0$ for any $j \in \mathcal{C}$. For these null control outcomes, their observed treatment effects, $\tau_{\mathcal{C}}^{\text{naive}}$, must equal the corresponding confounding bias. Since the bias is a function of the sensitivity vector β, we can establish the following constraint on β:

$$\tau_{\mathcal{C}}^{\text{naive}} = \frac{1}{\sigma_t^2\sqrt{1 - R^2_{T\sim U}}}\tilde{\Gamma}_{\mathcal{C}}\beta, \tag{5.35}$$

where $\tilde{\Gamma}_{\mathcal{C}}$ is a $c \times m$ matrix of $\tilde{\Gamma}$ that only contains rows corresponding to null controls. First, constraint 5.35 implies that $\tau_{\mathcal{C}}^{\text{naive}}$ must be in the column space of $\tilde{\Gamma}_{\mathcal{C}}$ so that the null control assumptions are compatible with the observed data; second, it also implies a lower bound on the magnitude of confoundedness, $R^2_{T\sim U}$, the fraction of confounding variation in the treatment. We formalize these ideas in the following proposition:

Proposition 5.4.1 *Suppose there are c known null control outcomes Y_j with $\tau_j = 0$ for $j \in \mathcal{C}$. Then, the null control compatibility condition $Q_{\tilde{\Gamma}_{\mathcal{C}}}\tau_j^{naive} = \tau_j^{naive}$ must hold, where $Q_{\tilde{\Gamma}_{\mathcal{C}}}$ denotes the projection matrix into the column space of $\tilde{\Gamma}_{\mathcal{C}}$. In addition, the fraction of variation in the treatment due to the confounding is lower bounded by*

$$R^2_{T\sim U} \geq R^2_{min} := \frac{\sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger}\tau_{\mathcal{C}}^{naive} \parallel_2^2}{1 + \sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger}\tau_{\mathcal{C}}^{naive} \parallel_2^2}, \tag{5.36}$$

where $\tilde{\Gamma}_{\mathcal{C}}^{\dagger}$ denotes a generalized inverse of $\tilde{\Gamma}_{\mathcal{C}}$.

Proof: See Appendix.

This lower bound quantifies the amount of confounding that becomes identifiable by null control assumptions. In the next theorem, we show that ignorance regions become smaller with null control assumptions.

Theorem 5.4.1 *For any value of* $\mathbb{R}^2_{T\sim U} \geq R^2_{\min}$ *which satisfies null control compatibility condition, the confounding bias for the treatment effect of the interested outcome* $a'Y$ *is in the interval*

$$a'\tilde{\Gamma}\tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{naive}_{\mathcal{C}} \pm \sqrt{\frac{R^2_{T\sim U}}{\sigma^2_t(1-R^2_{T\sim U})} - \parallel \tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{naive}_{\mathcal{C}} \parallel^2_2} \parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} \parallel_2, \tag{5.37}$$

where $P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}}$ *is the* $m \times m$ *projection matrix into the complement of the row space of* $\tilde{\Gamma}_{\mathcal{C}}$*.*
Proof: See Appendix.

Note that the ignorance region is no longer centered at $a'\tau^{\text{naive}}$ but instead $a'\tau^{\text{naive}} - a'\tilde{\Gamma}\tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{\text{naive}}_{\mathcal{C}}$ due to the additional information gained from the null controls. Also, Theorem 5.4.1 indicates that when $\tilde{\Gamma}_{\mathcal{C}}$ is full rank, treatment effects for *all* outcomes are identifiable. By comparing ignorance regions in Theorem 5.3.3 and 5.4.1, we can have the following corollary.

Corollary 5.4.1 *Under assumptions established in Theorem 5.4.1, null control outcomes reduce the width of the ignorance region by a multiplicative factor of*

$$\sqrt{1 - \frac{R^2_{min}}{1-R^2_{min}} \Big/ \frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}} \frac{\parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} \parallel_2}{\parallel a'\tilde{\Gamma} \parallel_2} \leq 1. \tag{5.38}$$

The above corollary indicates that null control reduce the width of the worst-case ignorance region in two ways. The first factor under the radical constrains the magnitude of unidentified confounding bias and reduces the width of the ignorance regions for all outcomes by an equal proportion. The second factor depends on the specific outcome of interest. The factor indicates that the ignorance region shrinks the most for outcomes that have the most similar confounder-outcome association with the null control outcomes. When $a'\tilde{\Gamma}$ is in the row space of $\tilde{\Gamma}_{\mathcal{C}}$, the treatment effect of $a'Y$ is identified.

When $a'\tilde{\Gamma}$ is orthogonal to the row space of $\tilde{\Gamma}_{\mathcal{C}}$, there is no further reduction of the ignorance region since $\frac{\|a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}}\|_2}{\|a'\tilde{\Gamma}\|_2} = 1$. In sum, the best null control outcomes are those which have large confounding bias and also have similar outcome-confounder association with that of the outcome of interest.

5.5 Analysis of Metabolomic Aging Clocks

In this section, we demonstrate our method in analysis of the effect of age on small molecules, called "metabolites", in humans. The metabolome consists of the structural and functional building blocks of an organism, which bridges genotype and phenotype and plays an important role in studies of aging and age-related traits [83]. We utilize targeted metabolomics dataset from [83] to investigate how aging affects the concentration of metabolites in cerebrospinal fluid (CSF). Inferring the biological effect of aging on the metabolome is complicated by potential confounders like by diet, exercise and lifestyle. For instance, younger people tend to exercise more than seniors, and exercise is known to significantly affect concentrations of some metabolites, such as lactate, pyruvate, TCA cycle intermediates [89].

Our dataset consists of 39 targeted metabolites measured for 85 individuals, and their corresponding age, ranging from 20 to 86 years old at time of healthy sample collection. We follow the data pre-processing used by [83] and use the R package `Amelia` to impute the missing values. We let T denote the age and $Y = (Y_1, \cdots, Y_q)(q = 39)$ be the measured concentration of targeted metabolites in CSF. Our estimand of interest is the biological effect of one year increase in age on each metabolite respectively. For simplicity, we model the outcome model with a linear regression, although other flexible outcome models like `BART` are also applicable [33]. With the linear assumption, our estimand reduces to the regression coefficient, τ_j, which is invariant to the specific level of t_1 and

t_2, i.e., our estimand is

$$\text{PATE}_{e_j} = \tau_j \tag{5.39}$$

for all $j = 1, \cdots, q$, and all t_1, t_2 such that $t_1 - t_2 = 1$.

We rescale all outcomes to unit variance and regress the scaled outcomes on age to estimate τ_j^{naive} under an assumption of no unobserved confounding, and we apply factor analysis using the R function `factanal` to the residual outcomes. We use cross validation to select the latent confounder dimension $m = 3$. First, we compute the ignorance regions of the treatment effects for all metabolites assuming $R^2_{T\sim U} \leq 95\%$ without incorporating any null control assumptions. As a result, all ignorance regions contain zero, which suggests that their treatment effects are sensitive to the unobserved confounding (see Figure B.1). In Table B.2.1 of the Appendix, we include robustness values for each metabolite, where the median is at 22% and the maximum is at 87% for glycerol 3-phosphate. In the literature of metabolomics studies, priori information is often used to assist analyses, for example, [90] utilizes metabolites that are known beforehand to be associated or unassociated with the biological factors of interest to determine whether their statistical approach for removing the unwanted variation has improved the analysis. Here, by making use of the null control outcomes, we show that we can further shrink the ignorance regions and make some of the metabolites with significant treatment effects distinguishable from the others.

Calibration with null Controls. Following the discussion in Section 5.4.2, we demonstrate the validity of our calibration method with null control outcomes. To demonstrate our approach, we use sorbitol, a sugar substitute, as a null control, as it has a reduced tendency to increase the sugar level in the blood and is used by diabetes patients and elderly individuals [91]. We plot the ignorance regions before and after the calibration with the null control, ordered by the extent to which they are influence by the null control cal-

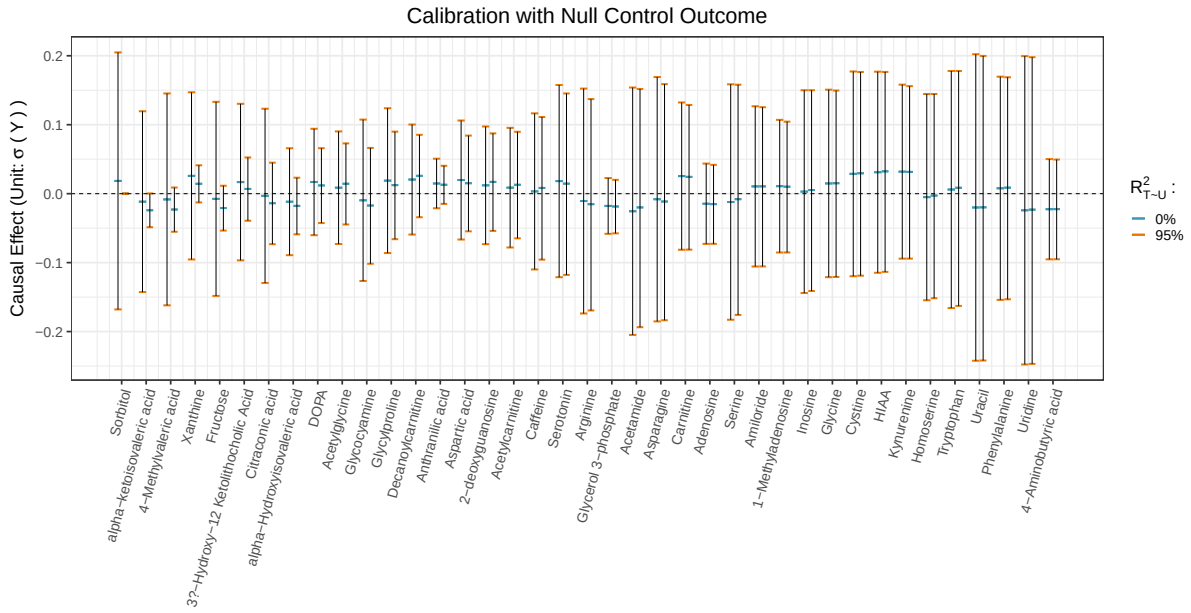


Figure 5.2: Estimated effect of increasing age by one year on abundances of metabolites before and after accounting for the null control outcome, sorbitol.

ibration (Figure 5.2). The ignorance region is constructed using 95% confidence interval for τ^{naive} in order to account for the estimation uncertainty of the observed data distribution. With the additional information about confounders provided by the null control, the ignorance region is largely reduced for treatment effects of the first few metabolites on the left in plot 5.2. Also, the estimate under the nonconfoundedness assumption (i.e. $R^2_{T\sim U} = 0$) for each metabolite changes after taking the null control outcome into account. Notably, we find that, after adjusting sorbitol to have a zero treatment effect, the effect of age on alpha-ketoisovaleric acid becomes robustly negative at level $R^2_{T\sim U} = 95\%$.

5.6 Discussion

In this chapter, we apply the copula-based sensitivity analysis to analyze multivariate outcomes with scalar treatments. Unlike previous work, which focuses on causal identification under strong assumptions, we explore the range of causal effects that are compat-

ible with the observed data under weaker assumptions about unobserved confounders. We show that bounds on the magnitude of confounding bias depend on the particular outcome of interest, and also provide practical exercises for calibrating the sensitivity parameters across multiple outcomes, and discuss the use of null control outcomes.

There are several directions we could further explore. First, the extension of our proposed method by relaxing the copula assumptions, as having been discussed in Section 4.8, would also be of great interest for the multi-outcome case here, including the generalization with variant conditional Gaussian copula that depends on t, or even with some non-Gaussian copula that characterize treatment-confounder relationship. In addition, we assume linearity of U in T to minimize the number of sensitivity parameters, ease calibration, and improve interpretability. More sophisticated calibration methods are needed for sensitivity parameters that characterize the nonlinear treatment-confounder associations. Second, to account for both the estimation uncertainty and unobserved confounding uncertainty, joint inference of the treatment and outcome models is preferable. This has been explored by [92] in the multi-treatment case, and we believe that it is also worth exploring in the multi-outcome case.

Either incorporating sparsity or null controls assumptions into the calibration of the sensitivity parameters can further constrain the partially identified regions in the multi-treatment or multi-outcome settings. [68] consider confounder adjustment in multiple hypothesis testing under the null controls and sparsity conditions respectively, where the null controls assumption requires the number of null control to be at least the dimension of the confounders' space while the sparsity condition requires at least half of the true treatment effects to be null. However, unlike the null controls assumption, there is no need for the sparsity assumption to pre-specify which treatment effects are likely to be null. As mentioned in [84], to achieve better performance, it's critical to choose null controls that are specially suited to the study.

Chapter 6

Discussion

In this dissertation, we explored several methods for assessing sensitivity to unobserved confounding in causal inference with structured and ordered treatments and outcomes.In all sensitivity analysis methods we proposed, the identifiable parameters are clearly separated from the unidentifiable parameters so that the observable predictions remain unperturbed by the sensitivity analysis. In Chapter 2, we introduced a reparameterization of the latent confounder model to explicitly decompose the effect of confounders on the outcome into confounding and non-confounding variation. In Chapter 3, we extended the Tukey's sensitivity analysis [2] to ordinal treatments with multiple levels. Finally, In Chapter 4 and 5, we focused on problems with multiple treatments and/or outcome using a copula-based framework. For all methods, we provide practical solutions to characterizing and calibrating the robustness of the causal treatment effects in these unique settings.

Although we have shown that our approaches work effectively under the corresponding settings, it would also be interesting to investigate how sensitive these methods are to violations of the underlying assumptions about the sensitivity analysis models. For instance, in the Tukey's approach, we may want to test whether our method would still

be valid if the true treatment assignment follows probit models instead the logits ones; or, in the copula-based approach, how the method would perform when the conditional association between Y, U given T actually follows some non-Gaussian coupla, such as Archimedean copulas, which may not be completely monotone.

There are several extensions and generalizations of these methods that could be explored. For example, in all our methods, we assume there are no interactions between confounders and treatments. To account for interactions, in the reparameterized latent confounder models, the sensitivity parameter which characterizes the conditional distribution of W given T should vary with the treatment. In the Tukey's method, the interaction would imply a more complicated treatment model beyond linearity in the logit scale. For the copula-based method, the interaction between U and T would suggest that the conditional copula, $c_{\psi_Y}(F_{Y|t}(y), F^{\psi_T}_{U|t}(u) \mid t)$, to be dependent on the multivariate treatment. As a consequence, the number of sensitivity parameters increases, which is a challenge for calibration and interpretation.

Lastly, we briefly explore a natural combination of the methods introduced in Chapter 4 and 5: can our method be applied to the setting in which both the treatment and outcome are multivariate? We demonstrated in model 4.11-4.13 that with multivariate treatments, the confounder-treatment relationship, parameterized by B, is identifiable up to a causal equivalence class, while the confounder-outcome relationship, parameterized by γ is unidentifiable. Conversely, in model 5.8-5.10 with multivariate outcomes, the confounder-outcome relationship, parameterized by Γ, is identifiable up to a causal equivalence class, while the confounder-treatment relationship, characterized by β, is unidentifiable. Consider the following linear model with multiple treatments and multi-

ple outcomes:

$$U = \epsilon_u, \quad \epsilon_u \sim N_m(0,\ \Sigma_u), \tag{6.1}$$

$$T = BU + \epsilon_{t|u}, \quad \epsilon_{t|u} \sim N_k(0,\ \sigma^2_{t|u} I_k), \tag{6.2}$$

$$Y = \mathcal{T}T + \Gamma U + \epsilon_{y|t,u}, \quad \epsilon_{y|t,u} \sim N_q(0,\ \sigma^2_{y|t,u} I_q),, \tag{6.3}$$

where B, $\mathcal{T}$ and Γ are respectively $k \times m$, $q \times k$ and $q \times m$ matrices. Note that, in the above model, the treatment assignment mechanism defined in Equation 6.2 is same as the one in model 4.11-4.13, where the conditional confounder distribution $f(u \mid t) \sim N(\mu_{u|t},\ \Sigma_{u|t})$ is identifiable, with $\mu_{u|t}$ and $\Sigma_{u|t}$ as known functions of B and $\sigma^2_{t|u}$. Additionally, as the outcomes are multivariate in Equation 6.3, the confounder-outcome relationship, characterized by Γ, can also be identified according to the discussion of Section 5.3.

Under model 6.1-6.3, the intervention distribution has density

$$f(y \mid do(T = t)) \sim N(\mathcal{T}T,\ \sigma^2_{y|t,u} I_q + \Gamma \Sigma_u \Gamma'), \tag{6.4}$$

and the observed outcome distribution can be written as

$$f(T \mid T = t) \sim N_q(\mathcal{T}^{\text{naive}} t,\ \Sigma_{y|t}), \tag{6.5}$$

where

$$\mathcal{T}^{\text{naive}} t = \mathcal{T}t + \Gamma \Sigma_u B' (B \Sigma_u B' + \sigma^2_{t|u} I_k)^{-1} t \tag{6.6}$$

$$= \mathcal{T}t + \Gamma \mu_{u|t}, \tag{6.7}$$

$$\Sigma_{y|t} = \sigma^2_{y|t,u} I_q + \Gamma (\Sigma_u - \Sigma_u B' (B \Sigma_u B' + \sigma^2_{t|u} I_k)^{-1} B \Sigma_u) \Gamma', \tag{6.8}$$

$$= \sigma^2_{y|t,u} I_q + \Gamma \Sigma_{u|t} \Gamma', \tag{6.9}$$

which are fully identifiable.

For given t_1, t_2, our estimand of interest, PATE_{t_1,t_2}, can be written as a function of

only the identifiable terms:

$$\text{PATE}_{t_1,t_2} = \mathcal{T}(t_1 - t_2), \tag{6.10}$$

$$= \mathcal{T}^{\text{naive}} t - \Gamma \mu_{u|t}, \tag{6.11}$$

which indicates that the treatment effects is identifiable when both treatments and outcomes are multivariate.

The situation we consider above is closely related to a line of work on identification with null control treatments and outcomes. [50] discusses conditions under which the average treatment effects can be nonparametrically identified with a null control treatment and a null control outcome, i.e. via a double null controls design, in the univariate treatment and outcome case. Further method developments in double null control design can be found in [93, 94]. Unlike previous works, we do not need to find external null controls as proxies of the unobserved confounder, but instead, we can learn information about the unobserved confounder from the multivariate correlation structures of the treatments and outcomes, and we show that it is sufficient for identifying treatment effects under model 6.1-6.3. There are many open questions about sensitivity analysis and identification with multivariate treatments and multivariate outcomes in more general models, which we leave to future work.

Appendix A

Appendix for Chapter 4

A.1 Theory

A.1.1 General Contrast Estimation Algorithm

Algorithm 3: Marginal Contrast Estimation for Arbitrary Copulas

Function `ComputeMean`(t, ψ)**:**
  for $k = 1, 2, \ldots, M$ **do**
    Sample y_k from $f(y \mid t)$;
    for $i = 1, 2, \ldots, n$ **do**
      Sample u_{ij} from $f(u \mid t_i)$;
      for $j = 1, 2, \ldots, N$ **do**
        Sample u_{ij} from $f(u \mid t_i)$;
        Compute $c_{ij} \leftarrow c_\psi(y_k, u_{ij} \mid t)$;
    Compute $w_k \leftarrow \frac{1}{nN} \sum_{ij} c_{ij}$;
  return $\frac{1}{M} \sum_k \nu(y_k) w_k$
return $\tau(\text{ComputeMean}(t_1, \psi), \text{ComputeMean}(t_2, \psi)$

A.1.2 Derivation of Algorithm 1

Since we have Equation 4.4 and 4.5, furthermore, we can write

$$E[v(Y) \mid do(t)] \ = \iint v(y) f(y \mid \tilde{y}) w_\psi(\tilde{y}, t) f(\tilde{y} \mid t) d\tilde{y} dy, \tag{A.1}$$

where $w_\psi(\tilde{y}, t) \approx \frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \left[\int c_\psi(F_{\tilde{Y}|t}(\tilde{y}), F_{U|t}(u) \mid t) f(u \mid t_i) du \right]$. To verify Algorithm 1, we only need to show that

$$\int f(\tilde{y} \mid t, u) f(u \mid t_i) du \sim N(\gamma^T(\mu_{u|t_i} - \mu_{u|t}),\ 1), \tag{A.2}$$

where $f(\tilde{y} \mid t, u) = f(\tilde{y} \mid t) c_\psi(F_{\tilde{Y}|t}(\tilde{y}), F_{U|t}(u) \mid t)$. According to Equation 4.7 and 4.8, we know that

$$f(u \mid t_i) \sim N(\mu_{u|t_i},\ \Sigma_{u|t}), \tag{A.3}$$

$$f(\tilde{y} \mid t, u) \sim N(\gamma^T(u - \mu_{u|t}),\ \sigma^2_{\tilde{y}|t,u}). \tag{A.4}$$

By integrating out the U, we have

$$\int f(\tilde{y} \mid t, u) f(u \mid t_i) du = \frac{1}{\sqrt{2\pi(\sigma^2_{\tilde{y}|t,u} + \gamma^T \Sigma_{u|t} \gamma)}} \exp\left\{ -\frac{(y - \gamma'(\mu_{u|t_i} - \mu_{u|t}))^2}{2(\sigma^2_{\tilde{y}|t,u} + \gamma^T \Sigma_{u|t} \gamma)} \right\}, \tag{A.5}$$

where $\sigma^2_{\tilde{y}|t,u} + \gamma' \Sigma_{u|t} \gamma = 1$.

A.1.3 Proof of Theorem 4.3.1

Theorem 4.3.1 *Assume model 4.7-4.9. Let* $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$ *where* $\mathcal{S}^+$ *is the space of symmetric positive definite matrices. Then* $[\psi_T]$ *is a causal equivalence class.*

Proof: The intervention distribution for $\tilde{y}$ is defined as

$$f_{\psi}(\tilde{y} \mid do(t)) = \int \left[\int f_{\psi_Y}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du \right] f(\tilde{t}) d\tilde{t} \tag{A.6}$$

where $\psi_Y = \gamma$ and $\psi_T = \{\mu_{u|t}, \Sigma_{u|t}\}$. Then, $\int f_{\gamma}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du \sim N(\gamma^T(\mu_{u|\tilde{t}} - \mu_{u|t}),\ 1)$ for any γ such that $\gamma' \Sigma_{u|t} \gamma \leq 1$ (see Equation A.5). Let $\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} \in [\psi_T]$ where $A \in \mathcal{S}^+$ is a positive definite matrix and assume $\tilde{\psi}_Y = \tilde{\gamma}$. Then,

$$\int f_{\tilde{\gamma}}(\tilde{y} \mid t, u) f_{\tilde{\psi}_T}(u \mid \tilde{t}) du \sim N(\tilde{\gamma}' A(\mu_{u|\tilde{t}} - \mu_{u|t}),\ 1). \tag{A.7}$$

Let $\tilde{\gamma} = A^{-1}\gamma$ be a bijective mapping from γ to $\tilde{\gamma}$. For any γ and positive definite A, we have $\tilde{\gamma}' A \Sigma_{u|t} A \tilde{\gamma} = \gamma' \Sigma_{u|t} \gamma \leq 1$ so that $\tilde{\gamma}$ is a valid copula parameter. In addition, $\int f_{\gamma}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du = \int f_{\tilde{\gamma}}(\tilde{y} \mid t, u) f_{\tilde{\psi}_T}(u \mid \tilde{t}) du$, which implies $f_{\gamma, \psi_T}(\tilde{y} \mid do(t)) = f_{\tilde{\gamma}, \tilde{\psi}_T}(\tilde{y} \mid do(t))$. Since Y is a deterministic function of $\tilde{Y}$, this implies $f_{\gamma, \psi_T}(y \mid do(t)) = f_{\tilde{\gamma}, \tilde{\psi}_T}(y \mid do(t))$. Therefore, $[\psi_T]$ is a causal equivalence class.

A.1.4 Proof of Theorem 4.4.1

Theorem 4.4.1 *Suppose that the observed data is generated by model 4.11-4.13. When there k treatments with $1 < m < k$, then ψ_T is identified up to the causal equivalence class $[\psi_T] = \{\tilde{\psi}_T = \{A\mu_{u|t}, A\Sigma_{u|t}A\} : A \in \mathcal{S}^+\}$. When there is a single treatment ($k = 1$) or $m = k$ confounders, then ψ_T is not identifiable up to causal equivalence class.*

Proof: The sample covariance matrix of the treatments is a consistent estimator for $B\Sigma_u B' + \Lambda_{t|u}$, the covariance matrix T. As long as $1 < m < k$, then the kth eigenvalue is a consistent estimator for $\sigma^2_{t|u}$. $B\Sigma_u B'$ is identified by the m eigenvectors and eigenvalues of the sample covariance matrix. The span of the first m eigenvectors of the sample

covariance matrix is a consistent estimator for the span of B. With model 4.11-4.13, the conditional distribution of confounder U

$$f(u \mid t) \sim N(\mu_{u|t},\ \Sigma_{u|t}), \tag{A.8}$$

where $\mu_{u|t} := \Sigma_u B'(B\Sigma_u B' + \Lambda_{t|u})^{-1}t$, $\Sigma_{u|t} := \Sigma_u - \Sigma_u B'(B\Sigma_u B' + \Lambda_{t|u})^{-1}B\Sigma_u$, and the intervention distribution

$$f_{\gamma,B,\Sigma_u}(y \mid do(T=t)) \sim N((\tau_{\text{naive}} - (B\Sigma_u B' + \Lambda_{t|u})^{-1}B\Sigma_u\gamma)'t,\ \sigma^2_{y|t,u} + \gamma'\Sigma_u\gamma), \tag{A.9}$$

where all m-vectors γ which satisfy $\gamma^T\Sigma_{u|t}\gamma \leq \sigma^2_{y|t}$ are valid sensitivity parameters.

Let $\tilde{B} = BA$ and $\tilde{\Sigma}_u = A^{-1}\Sigma_u A^{-T}$ for an arbitrary positive definite matrix A, so that $\tilde{B}\tilde{\Sigma}_u\tilde{B}' + \Lambda_{t|u} = B\Sigma_u B' + \Lambda_{t|u}$. Then, the observed treatments are consistent with $T = \tilde{B}\tilde{U} + \epsilon_{t|u}$, where $\tilde{U} = A^{-1}U \sim N_m(0, \tilde{\Sigma}_u)$. Hence, the conditional confounder distribution

$$f(\tilde{u} \mid t) \sim N_m(\tilde{\mu}_{u|t},\ \tilde{\Sigma}_{u|t}), \tag{A.10}$$

where $\tilde{\mu}_{u|t} = A^{-1}\mu_{u|t}$ and $\tilde{\Sigma}_{u|t} = A^{-1}\Sigma_{u|t}A^{-T}$. With $\tilde{B}$ and $\tilde{\Sigma}_u$, the intervention distribution can be alternatively expressed as

$$f_{\tilde{\gamma},\tilde{B},\tilde{\Sigma}_u}(y \mid do(T=t)) \sim N((\tau_{\text{naive}} - (\tilde{B}\tilde{\Sigma}_u\tilde{B}' + \sigma^2_{t|u}I_k)^{-1}\tilde{B}\tilde{\Sigma}_u\tilde{\gamma})'t,\ \sigma^2_{y|t,u} + \tilde{\gamma}'\tilde{\Sigma}_u\tilde{\gamma}) \tag{A.11}$$

with $\tilde{\gamma}$ satisfying $\tilde{\gamma}^T\tilde{\Sigma}_{u|t}\tilde{\gamma} \leq \sigma^2_{y|t}$ to be valid sensitivity parameter.

Let $\tilde{\gamma} = A^T\gamma$. If $\gamma^T\Sigma_{u|t}\gamma \leq \sigma^2_{y|t}$, then we have $\tilde{\gamma}^T\tilde{\Sigma}_{u|t}\tilde{\gamma} = \gamma^T\Sigma_{u|t}\gamma \leq \sigma^2_{y|t}$ and $f_{\tilde{\gamma},\tilde{B},\tilde{\Sigma}_u}(y \mid do(T=t)) = f_{\gamma,B,\Sigma_u}(y \mid do(T=t))$. Therefore, the causal equivalence class characterized by $\psi_T = \{\mu_{u|t}, \Sigma_{u|t}\}$ is identifiable when $1 < m < k$.

A.1.5 Proof of Theorem 4.4.2 and 4.4.3

Proof of Theorem 4.4.2

Theorem 4.4.2 *Suppose that the observed data is generated by model 4.11-4.13. Then, $\forall\gamma$ satisfying Assumptions 1 and 2,*

$$\gamma^T\Sigma_{u|t}\gamma \leq \sigma^2_{y|t}R^2_{Y\sim U|T}. \tag{A.12}$$

For any given Δt, we have

$$Bias^2_{\Delta t} \leq \sigma^2_{y|t}R^2_{Y\sim U|T} \parallel \Sigma^{-1/2}_{u|t}\mu_{u|\Delta t} \parallel^2_2 . \tag{A.13}$$

The bound is achieved when γ is colinear with $\Sigma^{-1}_{u|t}\mu_{u|\Delta t}$.

Proof: Under model 4.11-4.13, the variance of the observed outcome equals

$$\begin{aligned}\sigma^2_{y|t} &:= Var(Y \mid T)\\ &= \sigma^2_{y|t,u} + \gamma^T(I_m - B^T(BB^T + \Lambda_{t|u})^{-1}B)\gamma\\ &= \sigma^2_{y|t,u} + \gamma^T\Sigma_{u|t}\gamma,\end{aligned} \tag{A.14}$$

where $\gamma'\Sigma_{u|t}\gamma$ corresponds to the confounding variation, and $\sigma^2_{y|t,u}$ stands for the non-confounding variation in the residual of observed outcome. Hence, the fraction of confounding variation in the residual of Y, $R^2_{Y\sim U|T}$, can be expressed in terms of equation 4.25, which produces a constrain for γ (equation 4.26) that the confounding variation in the residual of Y, $\gamma^T\Sigma_{u|t}\gamma$, should not be larger than $\sigma^2_{y|t}R^2_{Y\sim U|T}$ for a given level of $R^2_{Y\sim U|T}$.

Let

$$Z := \Sigma^{1/2}_{u|t}\gamma, \tag{A.15}$$

then the omitted variable bias in equation 4.20 can be written as

$$Bias_{\Delta}t = Z^T \Sigma_{u|t}^{-1/2} \mu_{u|\Delta t},$$

where $Z^T Z \leq \sigma_{y|t}^2 R_{Y \sim U|T}^2$, implied by inequality 4.26.

Therefore,

$$\text{Bias}_{\Delta t}^2 = Z^T \Sigma_{u|t}^{-1/2} \mu_{u|\Delta t} \mu_{u|\Delta t}^T \Sigma_{u|t}^{-1/2} Z \tag{A.16}$$

$$\leq \sigma_{y|t}^2 R_{Y \sim U|T}^2 \parallel \Sigma_{u|t}^{-1/2} \mu_{u|\Delta t} \parallel_2^2, \tag{A.17}$$

where the bounds are reached when Z is colinear with $\Sigma_{u|t}^{-1/2} \mu_{u|\Delta t}$, i.e., γ is colinear with the $\Sigma_{u|t}^{-1} \mu_{u|\Delta t}$ inferred by the relationship defined in equation A.15.

Corollary 4.4.1 *Let d_1 be the largest singular value of B. For all Δt with $\parallel \Delta t \parallel_2 = 1$, the squared bias is bounded by*

$$Bias_{\Delta t}^2 \leq \frac{d_1^2}{(d_1^2 + \sigma_{t|u}^2)} \frac{\sigma_{y|t}^2}{\sigma_{t|u}^2} R_{Y \sim U|T}^2, \tag{A.18}$$

with equality when $\Delta t = u_1^B$, the first left singular vector of B. When $\Delta t \in Null(B')$, the naive estimate is unbiased, that is, $PATE_{\Delta t} = \tau'_{naive} \Delta t$.

Proof: Suppose that the matrix B has the singular value decomposition,

$$B = UDV^T,$$

where the diagonal entries of D are the singular values of B in descending order. Then, we can write

$$\mu_{u|\Delta t} = VD(D^2 + \sigma_{t|u}^2 I_s)^{-1} U^T \Delta t, \tag{A.19}$$

and

$$\Sigma_{u|t}^{-1} = V[I_s + \frac{1}{\sigma_{t|u}^2} D^2]V^T. \tag{A.20}$$

By plugging Equation A.19 and A.20 into the result of theorem 4.4.2, we have

$$\text{Bias}_{\Delta t}^2 \leq \frac{\sigma_{y|t}^2}{\sigma_{t|u}^2} R_{Y\sim U|T}^2 \parallel VD(\sigma_{t|u}^2 I_s + D^2)^{-1/2} U^T \Delta t \parallel_2^2, \tag{A.21}$$

where, according to Rayleigh quotient [95], the squared L2 norm reaches its maximum, $\frac{d_1^2}{(d_1^2+\sigma_{t|u}^2)}$, when Δt equals the first column of U, i.e., the first left singular vector of B.

Therefore, we have

$$\text{Bias}_{\Delta t}^2 \leq \frac{d_1^2}{(d_1^2 + \sigma_{t|u}^2)} \frac{\sigma_{y|t}^2}{\sigma_{t|u}^2} R_{Y\sim U|T}^2. \tag{A.22}$$

Proof of Theorem 4.4.3

Theorem 4.4.3 *Assume the model 4.7-4.9 with Gaussian outcomes. If $\Sigma_{u|t}$ is non-invertible, then $Bias_{t_1,t_2}$ is bounded if and only if $\mu_{u|t_1} - \mu_{u|t_2}$ is in the row space of $\Sigma_{u|t}$. When bounded,*
$Bias_{t_1,t_2}^2 \leq \sigma_{y|t}^2 R_{Y\sim U|T}^2 \|(\Sigma_{u|t}^{\dagger})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2})\|_2^2$, where $\Sigma_{u|t}^{\dagger}$ is the pseudo-inverse of $\Sigma_{u|t}$.

Proof: Under model 4.7-4.9, we have $\gamma^T \Sigma_{u|t} \gamma + \sigma_{\tilde{y}|t,u}^2 = 1$, where $\gamma^T \Sigma_{u|t} \gamma$ corresponds to the confounding variation and $\sigma_{\tilde{y}|t,u}^2$ corresponds to the non-confounding variation in the Gaussianized Y. Therefore, the confounding variation, $\gamma^T \Sigma_{u|t} \gamma$ should not be larger than a given level of $R_{\tilde{Y}\sim U|T}^2$,

$$\gamma^T \Sigma_{u|t} \gamma \leq R_{\tilde{Y}\sim U|T}^2 \tag{A.23}$$

where $R_{\tilde{Y}\sim U|T}^2$ denotes the fraction of confounding variation in residual variance of $\tilde{Y}$

conditional on T [1].

Let

$$Z := \Sigma_{u|t}^{1/2}\gamma, \tag{A.24}$$

then the omitted variable bias,

$$\text{Bias}_{t_1,t_2} = \sigma_{y|t} Z^T (\Sigma_{u|t}^{\dagger})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2}), \tag{A.25}$$

where $Z^T Z \leq R^2_{\tilde{Y}\sim U|T}$, implied by inequality A.23.

Therefore,

$$\begin{aligned}\text{Bias}^2_{t_1,t_2} &= \sigma^2_{y|t} Z^T (\Sigma_{u|t}^{\dagger})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2})(\mu_{u|t_1} - \mu_{u|t_2})^T (\Sigma_{u|t}^{\dagger})^{1/2} Z &\text{(A.26)}\\ &\leq \sigma^2_{y|t} R^2_{\tilde{Y}\sim U|T} \parallel (\Sigma_{u|t}^{\dagger})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2}) \parallel_2^2, &\text{(A.27)}\end{aligned}$$

where the bounds are reached when Z is colinear with $(\Sigma_{u|t}^{\dagger})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2})$, i.e., γ is colinear with the $\Sigma_{u|t}^{\dagger}(\mu_{u|t_1} - \mu_{u|t_2})$.

Suppose that $\Sigma_{u|t}$ has the eigendecomposition,

$$\Sigma_{u|t} = Q\Lambda Q^T, \tag{A.28}$$

where Q is the square $s \times s$ matrix whose jth column is the eigenvector q_j of $\Sigma_{u|t}$, and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{jj} = \lambda_j$, in descending order. If $\Sigma_{u|t}$ is non-invertible and has rank p ($p \leq s$), then we have $\lambda_j = 0$ for $j = p+1, \cdots, s$.

On the one hand, when $\mu_{u|t_1} - \mu_{u|t_2}$ is in the row space of $\Sigma_{u|t}$, it can be expressed as a linear combination of q_j, $\sum_{j=1}^{p} a_j q_j$, $a_j \in \mathbb{R}$. Then, we have the squared omitted

[1] $R^2_{\tilde{Y}\sim U|T}$ coincides with $R^2_{\tilde{Y}\sim U}$ here, but we use notation $R^2_{\tilde{Y}\sim U|T}$ for consistency.

variable bias

$$\text{Bias}^2_{t_1,t_2} \leq \sigma^2_{y|t} R^2_{\tilde{Y}\sim U|T} \parallel (\Sigma^{\dagger}_{u|t})^{1/2}(\mu_{u|t_1} - \mu_{u|t_2}) \parallel^2_2, \tag{A.29}$$

$$= \sigma^2_{y|t} R^2_{\tilde{Y}\sim U|T} \parallel Q(\Lambda^{\dagger})^{1/2} Q^T \sum_{j=1}^{p} a_j q_j \parallel^2_2, \tag{A.30}$$

$$= \sigma^2_{y|t} R^2_{\tilde{Y}\sim U|T} \sum_{i=1}^{s} (\sum_{j=1}^{p} a_j \lambda_j^{-\frac{1}{2}} Q_{ij})^2, \tag{A.31}$$

where Q_{ij} denotes the element at the ith row and jth column of matrix Q, and $\Lambda^{\dagger}$ is the pseudo-inverse of Λ by taking the reciprocal of each its non-zero element on the diagonal, leaving the zeros in place.

On the other hand, when $\text{Bias}^2_{t_1,t_2}$ is bounded, let's assume that $\mu_{u|t_1} - \mu_{u|t_2}$ is not in the row space of $\Sigma_{u|t}$, say $\mu_{u|t_1} - \mu_{u|t_2} = q_s$. Since $\lambda_s = 0$, $\lambda_s^{-1/2} = \infty$ so as the bound of $\text{Bias}^2_{t_1,t_2}$ equal to ∞, which contradicts the condition that $\text{Bias}^2_{t_1,t_2}$ is bounded.

Therefore, $\text{Bias}^2_{t_1,t_2}$ is bounded if and only if $\mu_{u|t_1} - \mu_{u|t_2}$ is in the row space of $\Sigma_{u|t}$.

A.1.6 Proof of Proposition 4.5.1, Theorem 4.5.1 and Corollary

Proof of Proposition 4.5.1

Proposition 4.5.1 *Suppose there are c known null control treatment contrasts, t_{j1} versus t_{j2} for $j \in \mathcal{C}$. Then, the null control compatibility condition $\mu_{y|\Delta t_c} P_{M_{u|\Delta t_c}} = \mu_{y|\Delta t_c}$ must hold where $P_{M_{u|\Delta t_c}}$ denotes the projection matrix into the row space of $M_{u|\Delta t_c}$. Additionally, the partial fraction of variance explained due to confounders given treatments is lower bounded by*

$$R^2_{Y\sim U|T} \geq R^2_{min} = \frac{1}{\sigma^2_{y|t}} \parallel \mu_{y|\Delta t_c} M^{\dagger}_{u|\Delta t_{\mathcal{C}}} \parallel^2_2 .$$

where $M^{\dagger}_{u|\Delta t_{\mathcal{C}}}$ *denotes a generalized inverse of* $M_{u|\Delta t_{\mathcal{C}}}$.

Proof: Assume there are c null control treatment contrasts, satisfying

$$\sigma_{y|t}\gamma'\Sigma^{1/2}_{u|t}M_{u|\Delta t_{\mathcal{C}}} = \mu_{y|\Delta t_{\mathcal{C}}} \tag{A.32}$$

The solution for above equation exists if and only if $\mu_{y|\Delta t_{\mathcal{C}}}M^{\dagger}_{u|\Delta t_c}M_{u|\Delta t_{\mathcal{C}}} = \mu_{y|\Delta t_{\mathcal{C}}}$ holds, which ensures that the null control assumptions are compatible. Under this condition, all solutions to equation A.32 can be expressed as

$$\gamma' = \frac{1}{\sigma_{y|t}}\mu_{y|\Delta t_{\mathcal{C}}}M^{\dagger}_{u|\Delta t_c}(\Sigma^{\dagger}_{u|t})^{1/2} + w'(I - M_{u|\Delta t_c}M^{\dagger}_{u|\Delta t_c})(\Sigma^{\dagger}_{u|t})^{1/2}, \tag{A.33}$$

Since $\gamma'\Sigma_{u|t}\gamma = R^2_{Y\sim U|T}$, w can be any $m \times 1$ vector satisfying

$$\parallel w'(I - M_{u|\Delta t_c}M^{\dagger}_{u|\Delta t_c}) \parallel^2_2 = R^2_{Y\sim U|T} - \frac{1}{\sigma^2_{y|t}} \parallel \mu_{y|\Delta t_{\mathcal{C}}}M^{\dagger}_{u|\Delta t_c} \parallel^2_2$$

Further,

$$\parallel w'(I - M_{u|\Delta t_c}M^{\dagger}_{u|\Delta t_c}) \parallel^2_2 \geq 0$$

must hold, we know that $R^2_{Y\sim U|T}$ must be at least

$$\frac{1}{\sigma^2_{y|t}} \parallel \mu_{y|\Delta t_{\mathcal{C}}}M^{\dagger}_{u|\Delta t_c} \parallel^2_2,$$

which proves Proposition 4.5.1.

Proof of Theorem 4.5.1 and Corollary

Theorem 4.5.1 *For any value of* $R^2_{Y\sim U|T} > R^2_{\min}$ *which satisfies null control compatibility condition, the confounding bias for the treatment effect of contrast* Δt *is in the*

interval

$$\mu_{y|\Delta t_c} M^{\dagger}_{u|\Delta t_{\mathcal{C}}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \pm \tag{A.34}$$

$$\sigma_{y|t} \sqrt{R^2_{Y\sim U|T} - R^2_{\min}} \left\| Q^{\perp}_{M_{u|\Delta t_c}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \right\|_2 \tag{A.35}$$

where $Q^{\perp}_{M_{u|\Delta t_c}}$ *is the* $m \times m$ *projection matrix into the complement of the column space of* $M_{u|\Delta t_c}$.

Corollary *Under the assumptions established in Theorem 1, null controls reduce the width of the partial identification ignorance region by a multiplicative factor of*

$$\sqrt{1 - R^2_{\min}/R^2_{Y\sim U|T}} \frac{\| \, Q^{\perp}_{M_{u|\Delta t_c}} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \, \|_2}{\| \, (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \, \|_2} \leq 1 \tag{A.36}$$

Proof: For treatment contrast, t_1 versus t_2, the omitted variable bias of PATE_{t_1,t_2} is $\text{Bias}_{t_1,t_2} = \gamma'(\mu_{u|t_1} - \mu_{u|t_2})$ and so

$$\mu_{y|\Delta t_{\mathcal{C}}} M^{\dagger}_{u|\Delta t_c} (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \pm \tag{A.37}$$

$$\sigma_{y|t} \sqrt{R^2_{Y\sim U|T} - R^2_{\min}} \left\| (I - M_{u|\Delta t_c} M^{\dagger}_{u|\Delta t_c}) (\Sigma^{\dagger}_{u|t})^{1/2} (\mu_{u|t_1} - \mu_{u|t_2}) \right\|_2,$$

where the bounds are achieved when $(I - M_{u|\Delta t_c} M^{\dagger}_{u|\Delta t_c}) w$ has the largest cosine similarity with $\Sigma^{-1/2}_{u|t} \mu_{u|\Delta t}$.

Compare the second term of A.37 and the bound given in Theorem 4.4.3, we see that

the width of ignorance region is shrunk by a multiplicative factor of

$$\sqrt{1 - R^2_{\min}/R^2_{Y\sim U|T}} \frac{\| \ (I - M_{u|\Delta t_c} M^{\dagger}_{u|\Delta t_c})(\Sigma^{\dagger}_{u|t})^{1/2} \mu_{u|\Delta t} \ \|_2}{\| \ (\Sigma^{\dagger}_{u|t})^{1/2} \mu_{u|\Delta t} \ \|_2}. \tag{A.38}$$

A.2 Modeling Choice Details

A.2.1 Identification and Inference in the Factor Model

Here, we briefly elaborate on identifiability of the probabilistic principal components model, which is a prerequisite for our multi-cause sensitivity analysis. Identifiability under various factor model assumptions is well studied and has a long history in the literature [96, 97]. In the specific probabilistic principal components model 4.12, Tipping and Bishop (1999) [98] provide a maximum likelihood solution for inferring the latent confounder parameters conditional on m. Many procedures are available for selecting the appropriate value of m, using for example Bayesian model selection techniques [99] or large p, small n asymptotics Gavish and Donoho (2014) [100].

The change of variables described at the end of the subsection 4.4.1 further elucidates important situations in which we cannot bound the omitted variable bias due to non-identifiability of the factor model. Again, we focus on the rotated treatments $\tilde{T} \sim N(0, \Delta + \sigma^2_{t|u} I_k)$ and highlight two simple situations in which we cannot bound the the causal effects. First, when B is rank k, i.e. there exist $m = k$ independent confounders, Δ has no non-zero entries on the diagonal and thus we cannot identify either Δ nor $\text{Cov}(\epsilon_{t|u}) = \sigma^2_{t|u} I_k$, only their sum. Second, if $\text{Cov}(\epsilon_{t|u})$ is an unknown arbitrary diagonal matrix (as opposed to a matrix proportional to the identity), then $\text{Cov}(\epsilon_{t|u})$ is

not distinguishable from Δ. In both of these cases, the worst-case bias is unbounded since the non-confounding variation of the treatment assignment, $\text{Cov}(\epsilon_{t|u})$, can be arbitrarily small. In such settings, we can still apply approaches used in single cause sensitivity analysis, by specifying both Ψ_Y and Ψ_T; when the factor model is not identifiable, Ψ_T must be chosen as a true parameter, e.g. by bounding the fraction of treatment variation due to confounding, $R^2_{T\sim U}$.

A.2.2 Confounder Inference with Variational Autoencoders

Probabilistic Principal Component Analysis should only be used when the treatments are approximately Gaussian treatments. For binary and other general treatment distributions, more sophisticated probabilistic latent variables models are required. Examples of such latent variable models include models for count data like the logistic factor analysis [101] and Poisson factor analysis methods [102]. Unfortunately, these models imply posteriors which are non-Gaussian and heteroskedastic, violating Assumptions 4.3.2 and 4.3.1.

As such, for general treatment distributions, our approach is to infer a conditional Gaussian latent variable model using a variational autoencoder (VAE). VAEs have been extremely popular in machine learning, in particular for generating low dimensional representations of complex inputs like images [103] but more recently have been used in scientific and decision-making applications [104] and in applications to causal inference [105]. A VAE consists of a prior distribution, $f(u)$, typically for the low-dimensional latent variables, a stochastic encoder, and a stochastic decoder. In our application, the inferred stochastic decoder, $\hat{f}_\theta(t \mid u)$, is a non-linear map from latent confounders to a distribution over causes. Together, the prior distribution for u and the decoder imply a posterior confounder distribution, $\hat{f}(u \mid t)$.

In practice, inference for the true posterior is intractable and so a variational approximation, called the encoder, $q_\phi(u \mid t)$, is used in place of the true posterior. Typically the encoder is chosen to be a normal distribution with mean and variance which are non-linear functions of the input, $q_\phi = N(\mu_\phi(t), \sigma^2_\phi(t))$. A crucial question is that how well the Gaussian encoder approximates the true posterior; improving the variational approximation to the true latent variable posterior is an area of active research. In this work, we follow a common strategy of using the encoder learned by the VAE as the proposal distribution in an importance sampler [104].

Specifically, we apply a variant of the Constant-Variance Variational Autoencoder (CV-VAE) [106] to infer the conditional confounder distribution, $f(u \mid t) \sim N(\mu_{u|t}, \Sigma_{u|t})$, in which $\Sigma_{u|t}$ does not depend on the level of t. We use the importance sampling to improve estimates of the conditional mean $\mu_{u|t}$, and posterior variance, $\Sigma_{u|t}$. While this appraoch only yields an approximation to the true posterior, we demonstrate the practical effectiveness of this approach in Sections 4.6 and 4.7.

A.2.3 Binary Outcomes

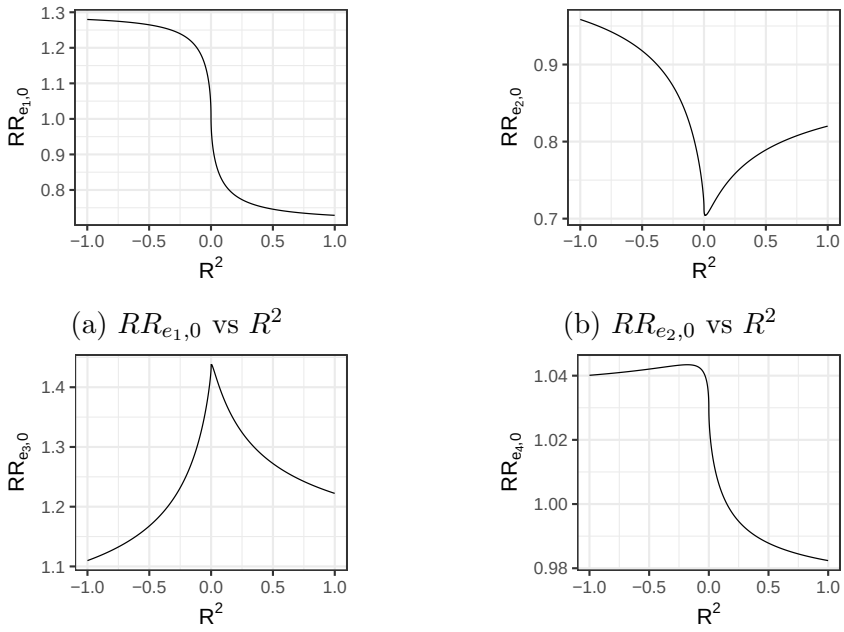
For binary outcomes with the risk ratio estimand:

$$RR_{t,\bullet} = \sum_{t_i \in \mathcal{T}} \Phi\left(\Phi^{-1}(\mu_{y|t}) + \gamma^T(\mu_{u|t_i} - \mu_{u|t})\right) \Big/ Pr(Y=1), \tag{A.39}$$

which implies that

$$RR_{t_1,t_2} = \sum_{t_i \in \mathcal{T}} \Phi\left(\Phi^{-1}(\mu_{y|t_1}) + \gamma^T(\mu_{u|t_i} - \mu_{u|t_1})\right) \Big/ \sum_{t_i \in \mathcal{T}} \Phi\left(\Phi^{-1}(\mu_{y|t_2}) + \gamma^T(\mu_{u|t_i} - \mu_{u|t_2})\right), \tag{A.40}$$

where $\gamma^T \Sigma_{u|t} \gamma \leq \sigma^2_{\tilde{y}|t} R^2_{\tilde{Y} \sim U|T}$. We can numerically explore values of RR_{t_1,t_2} within the valid domain of γ, and calculate the corresponding implicit partial R-squared by $R^2_{\tilde{Y} \sim U|T} = \frac{\gamma^T \Sigma_{u|t} \gamma}{\sigma^2_{\tilde{y}|t}}$. To calculate the robustness value, we only need to find the value of $R^2_{\tilde{Y} \sim U|T}$ for which the corresponding $RR_{t_1,t_2} = 1$. Noticeably, RR_{t_1,t_2} is not monotone in $R^2_{\tilde{Y} \sim U|T}$, since the variance of intervention distribution also depends on γ. This is evident in the simulation in Section 4.6 where we fit the observed outcome model by probit regression and the valid range for scalar γ is $[-\frac{1}{\sigma_{u|t}}, \frac{1}{\sigma_{u|t}}]$. We visualize the non-monotone relationship between RR_{t_1,t_2} and $R^2_{\tilde{Y} \sim U|T}$ in Figure A.1.



(a) $RR_{e_1,0}$ vs R^2

(b) $RR_{e_2,0}$ vs R^2

(c) $RR_{e_3,0}$ vs R^2

(d) $RR_{e_4,0}$ vs R^2

Figure A.1: RR_{t_1,t_2} is non-monotone in $R^2_{\tilde{Y}\sim U|T}$. Positive values of R^2 indicates that U is positively correlated with $\tilde{Y}$, and negative values of R^2 means that U is negatively correlated with $\tilde{Y}$.

A.3 Additional Results

A.3.1 Additional Results from Simulation in Sparse Effects Setting

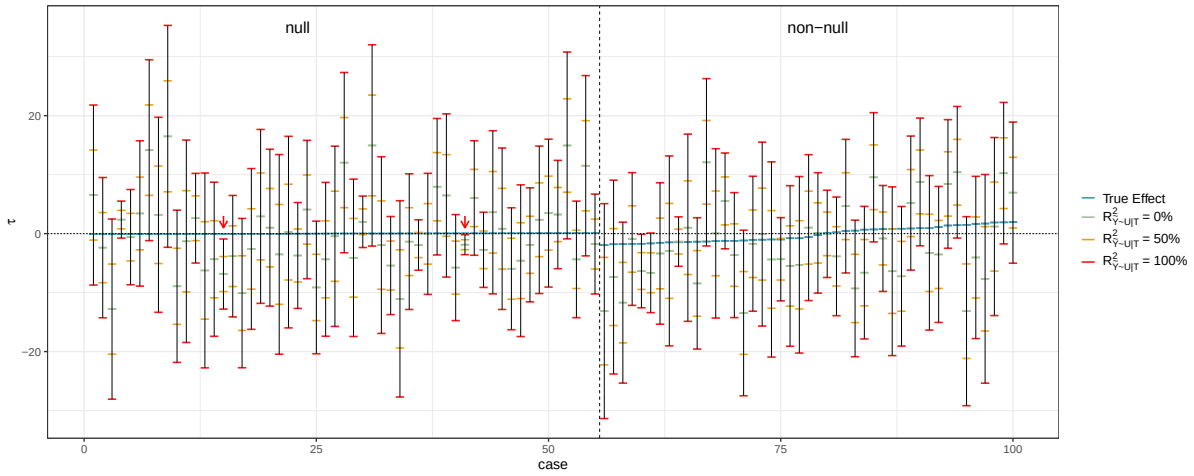


Figure A.2: Worst-case ignorance regions for 55 randomly chosen null effects (left) and all 45 non-null effects (right) ordered by the magnitude of true effects in each group. Two red arrows indicate non-null treatments for which the worst-case ignorance region does not cover the true effect. This appears to be due to estimation error in the outcome model, more so than with the VAE.

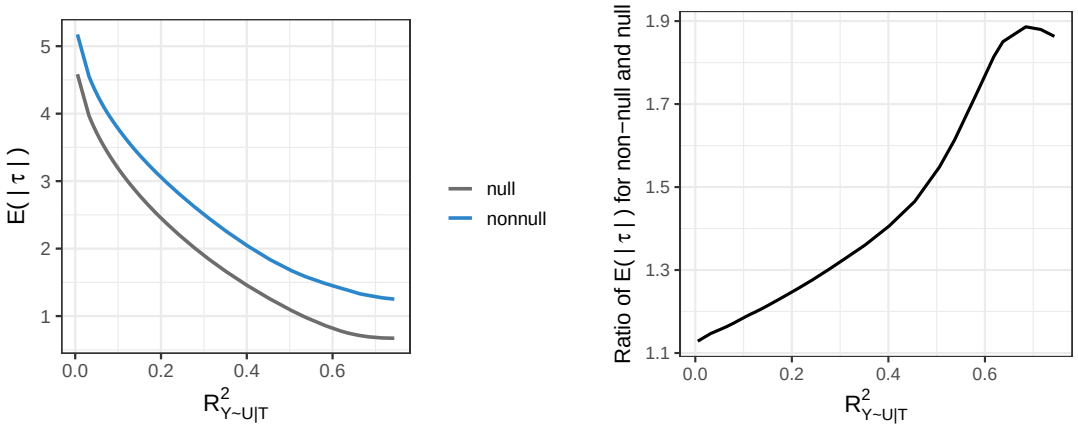
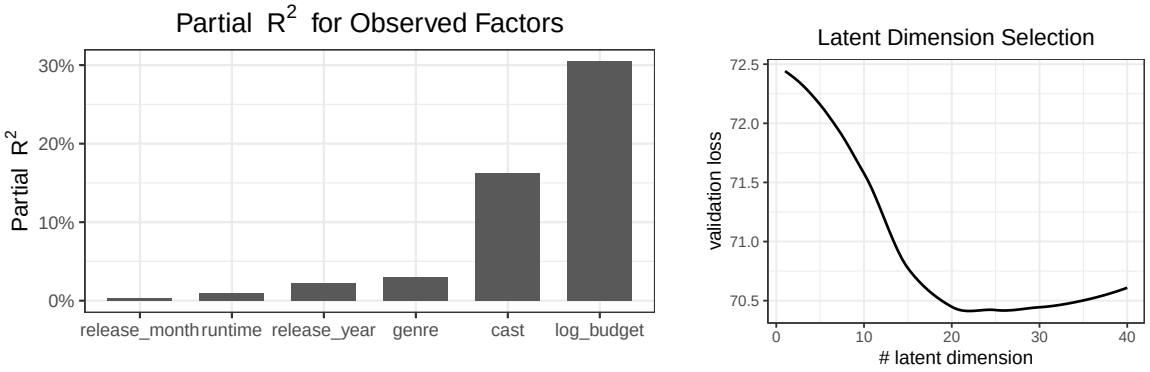


Figure A.3: Change in $E(|\tau|)$ for the L1-minimized estimates as a function of $R^2_{Y\sim U|T}$, separated by null and non-null effects. (a) The magnitude of effects decreases with $\mathcal{R}^2$, with a larger relative decrease for null contrasts. (b) The relative magnitude of non-null and null effects increases with $\mathcal{R}^2$ in general. The magnitude of non-null effects can be as large as 1.9 times the null effects when $\mathcal{R}^2$ is large.

A.3.2 Additional Results from the Actor Case Study



(a) (b)

Figure A.4: (a) Estimated partial R^2 for observed confounders using method described in section 4.5.1. Budget is the most dominant variable, which can explain significantly higher variation in outcome Y. (b) Latent confounder dimension selection, based on the reconstruction loss on the validation set.

Table A.1: Robustness Value for Significant Actors

	Effect	$RV_{mean}(\%)$	$RV_{limit}(\%)$
John Ratzenberger	52.91	21.79	9.36
Tom Cruise	49.48	5.09	1.59
Stan Lee	47.53	14.43	4.16
Morgan Freeman	41.95	1.63	0.26
Will Smith	41.87	8.61	2.64
Bruce Willis	41.44	1.86	0.3
Tom Hanks	36.40	4.16	0.79
Harrison Ford	35.26	3.25	0.6
Arnold Schwarzenegger	34.13	3.39	0.55
Johnny Depp	31.53	1.27	0.08
Frank Welker	30.99	2.86	0.34
Brad Pitt	30.61	1.54	0.09
Judi Dench	30.01	8.87	1.41
Leonardo DiCaprio	29.52	36.47	7.35
Adam Sandler	28.39	3.22	0.19
Liam Neeson	27.40	0.91	0.03
Denzel Washington	26.09	2.01	0.11
Eddie Murphy	25.82	3.30	0.27
John Travolta	23.96	2.45	0.13
Robin Williams	23.26	1.12	0.03
Hugo Weaving	23.02	4.20	0.12
Michael Caine	22.23	3.03	0.11
Channing Tatum	22.22	2.33	0.06
Angelina Jolie	20.81	3.72	0.09
Carla Gugino	20.74	4.24	0.13
Octavia Spencer	20.68	5.78	0.13
Kathy Bates	20.58	8.25	0.17
Jim Carrey	19.79	2.79	0
Ian McKellen	19.02	4.70	0
Reese Witherspoon	18.97	8.17	0.27
Zoe Saldana	18.79	7.33	0.12
Kevin Hart	18.41	3.21	0.05
Timothy Spall	18.40	9.39	0.25
Jamie Foxx	18.15	6.02	0.01
Judy Greer	17.76	4.17	0.01
Tommy Lee Jones	17.67	2.96	0
Rose Byrne	16.85	7.62	0.05
Andy Serkis	16.53	9.48	0.05
Dennis Hopper	-14.79	19.32	0
Kate Bosworth	-16.01	5.20	0.04
Viggo Mortensen	-16.24	6.50	0.02
Tim Blake Nelson	-16.58	24.78	0.09
Jeremy Piven	-17.45	4.86	0.06
Elias Koteas	-21.64	13.87	1.06
Susan Sarandon	-22.57	7.90	0.24
Mark Ruffalo	-23.29	3.17	0.11

Appendix B

Appendix for Chapter 5

B.1 Theory

B.1.1 Proof of Theorem 5.3.1

Theorem 5.3.1 *Assume model 5.9-5.6. Let* $[\psi_Y] = \{\tilde{\psi}_Y = \{\Gamma A\} : A \in \mathcal{S}^+\}$ *where* $\mathcal{S}^+$ *is the space of symmetric positive definite matrices. Then* $[\psi_Y]$ *is a causal equivalence class.*

Proof: The intervention distribution for $\tilde{y}$ is defined as

$$f_\psi(\tilde{y} \mid do(t)) = \int \left[\int f_{\psi_Y}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du \right] f(\tilde{t}) d\tilde{t}. \tag{B.1}$$

Under model 5.9-5.6, we have

$$f_{\psi_Y}(\tilde{y} \mid t, u) \sim N_q(\Gamma(u - \mu_{u|t}),\ \Lambda_{\tilde{y}|t,u}), \tag{B.2}$$

$$f_{\psi_T}(u \mid t) \sim N(\mu_{u|t},\ \Sigma_{u|t}), \tag{B.3}$$

with $\mu_{u|t} = \frac{\Sigma_u \beta}{\sigma_t^2}(t - \mu_t)$, $\Sigma_{u|t} = \Sigma_u - \frac{\Sigma_u \beta \beta' \Sigma_u}{\sigma_t^2}$, and parameters $\psi_T = \beta$, $\psi_Y = \Gamma$.

Thus, it can be derived that

$$\int f_{\psi_Y}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du \sim N_q(\frac{\Gamma \Sigma_u \beta}{\sigma_t^2}(\tilde{t} - t),\ \Lambda_{\tilde{y}|t,u} + \Gamma \Sigma_{u|t} \Gamma^T) \tag{B.4}$$

with sensitivity parameter β satisfying $\beta' \Sigma_u \beta \leq \sigma_t^2$.

Let $\tilde{\psi}_Y = \Gamma A \in [\psi_Y]$ and $\tilde{U} = A^{-1}U$, with A being a symmetric positive definite matrices, such that $\tilde{\Gamma} \tilde{\Sigma}_{u|t} \tilde{\Gamma}' + \Lambda_{\tilde{y}|t,u} = \Gamma \Sigma_{u|t} \Gamma' + \Lambda_{\tilde{y}|t,u} = C_{y|t}$, where $\tilde{\Sigma}_u := \text{Cov}(\tilde{U}) = \text{Cov}(A^{-1}U) = A^{-1} \Sigma_u A^{-T}$ and $\tilde{\Sigma}_{u|t} := \text{Cov}(\tilde{U} \mid t) = \text{Cov}(A^{-1}U \mid t) = A^{-1} \Sigma_{u|t} A^{-T}$. Assume that $\tilde{\psi}_T = \tilde{\beta}$. With $\tilde{\psi}_T$, $\tilde{\psi}_Y$ and $\tilde{U}$, the distribution in Equation B.4 can be alternatively expressed as

$$\int f_{\tilde{\psi}_Y}(\tilde{y} \mid t, u) f_{\tilde{\psi}_T}(u \mid \tilde{t}) du \sim N_q(\frac{\tilde{\Gamma} \tilde{\Sigma}_u \tilde{\beta}}{\sigma_t^2}(\tilde{t} - t),\ \Lambda_{\tilde{y}|t,u} + \tilde{\Gamma} \tilde{\Sigma}_{u|t} \tilde{\Gamma}^T). \tag{B.5}$$

Let $\tilde{\beta} = A^T \beta$ be a bijective mapping from β to $\tilde{\beta}$. For any β and rotation matrix A, we have $\tilde{\beta}' \tilde{\Sigma}_u \tilde{\beta} = \beta' \Sigma_u \beta \leq \sigma_t^2$ so that $\tilde{\beta}$ is a valid sensitivity parameter, and it's easy to see that $\int f_{\psi_Y}(\tilde{y} \mid t, u) f_{\psi_T}(u \mid \tilde{t}) du = \int f_{\tilde{\psi}_Y}(\tilde{y} \mid t, u) f_{\tilde{\psi}_T}(u \mid \tilde{t}) du$, which implies that $f_{\psi_Y, \psi_T}(\tilde{y} \mid do(t)) = f_{\tilde{\psi}_Y, \tilde{\psi}_T}(\tilde{y} \mid do(t))$. Since Y is a deterministic function of $\tilde{Y}$, this implies $f_{\psi_T, \psi_Y}(y \mid do(t)) = f_{\tilde{\psi}_T, \tilde{\psi}_Y}(y \mid do(t))$. Therefore, $[\psi_Y]$ is a causal equivalence class.

B.1.2 Proof of Theorem 5.3.2

Theorem 5.3.2 *Suppose that the observed data is generated by model 5.8-5.10. When there q outcomes with $1 < m < q$, then ψ_Y is identified up to the causal equivalence class $[\psi_Y] = \{\tilde{\psi}_Y = \{\Gamma A\} : A \in \mathcal{S}^+ \}$. When there is a single outcome ($q = 1$) or at least $m = q$ confounders, then ψ_Y is not identifiable up to causal equivalence class.*

Proof: Under model 5.8-5.10, the intervention distribution has density

$$f_{\beta,\Gamma,\Sigma_u}(y \mid do(T=t)) \sim N((\tau^{\text{naive}} - \frac{\Gamma\Sigma_u\beta}{\sigma_t^2})t,\ \Lambda_{y|t,u} + \Gamma\Sigma_u\Gamma'), \tag{B.6}$$

where $\sigma_t^2 := \beta'\Sigma_u\beta + \sigma_{t|u}^2$, denoting the marginal variance of treatment, and all m-vectors β which satisfy $\beta'\Sigma_u\beta \leq \sigma_t^2$ are valid sensitivity parameters.

Let $\tilde{\Gamma} = \Gamma A$, $\tilde{U} = A^{-1}U$. Then, we have $\tilde{\Sigma}_u := \text{Cov}(\tilde{U}) = A^{-1}\text{Cov}(U)A^{-T} = A^{-1}\Sigma_u A^{-T}$ and $\tilde{\Sigma}_{u|t} := \text{Cov}(\tilde{U} \mid t) = A^{-1}\text{Cov}(U \mid t)A^{-T} = A^{-1}\Sigma_{u|t}A^{-T}$, and thus $\text{Cov}(Y \mid t) = \Gamma\Sigma_{u|t}\Gamma' + \Lambda_{y|t,u} = \tilde{\Gamma}\tilde{\Sigma}_{u|t}\tilde{\Gamma}' + \Lambda_{y|t,u}$, which are both compatible with the observed data. With $\tilde{\Gamma}$ and $\tilde{\Sigma}_u$, the intervention distribution can be alternatively expressed as

$$f_{\tilde{\beta},\tilde{\Gamma},\tilde{\Sigma}_u}(y \mid do(T=t)) \sim N((\tau^{\text{naive}} - \frac{\tilde{\Gamma}\tilde{\Sigma}_u\tilde{\beta}}{\sigma_t^2})t,\ \Lambda_{y|t,u} + \tilde{\Gamma}\tilde{\Sigma}_u\tilde{\Gamma}') \tag{B.7}$$

for any valid sensitivity parameter $\tilde{\beta}$ satisfying $\tilde{\beta}'\tilde{\Sigma}_u\tilde{\beta} \leq \sigma_t^2$.

Let $\tilde{\beta} = A^T\beta$. If $\beta'\Sigma_u\beta \leq \sigma_t^2$, then we have $\tilde{\beta}'\tilde{\Sigma}_u\tilde{\beta} = \beta'\Sigma_u\beta \leq \sigma_t^2$, and it can be easily seen that $f_{\tilde{\beta},\tilde{\Gamma},\tilde{\Sigma}_u}(y \mid do(T=t)) = f_{\beta,\Gamma,\Sigma_u}(y \mid do(T=t))$. Therefore, the causal equivalence class characterized by $\psi_Y = \{\Gamma\}$ is identifiable when $1 < m < q$.

B.1.3 Proof of Theorem 5.3.3

Theorem 5.3.3 *Suppose that the observed data is generated by model 4.11-4.13. Then,* $\forall\beta$ *satisfying*

$$\beta'\Sigma_u\beta = \sigma_t^2 R_{T\sim U}^2 \tag{B.8}$$

with $0 \leq R_{T\sim U}^2 < 1$. *For given* a, *the confounding bias is bounded by*

$$Bias_a^2 \leq \frac{1}{\sigma_t^2}\frac{R_{T\sim U}^2}{1 - R_{T\sim U}^2} \parallel a'\tilde{\Gamma} \parallel_2^2, \tag{B.9}$$

where the bound is achieved when β is colinear with $a'\tilde{\Gamma}$.

Proof: The sensitivity parameter β can be reparameterized in terms of a direction d^β and an R-squared:

$$\beta = d^\beta u^\beta, \tag{B.10}$$

where $d^\beta = \sigma_t\sqrt{R^2_{T\sim U}}$ and $u^\beta \in \mathcal{C}^{m-1}$ is a m-dimensional unit vector.

Therefore, we can write the eigendecomposition of matrix $I_m - \frac{\beta\beta'}{\sigma_t^2}$ as

$$I_m - \frac{\beta\beta'}{\sigma_t^2} = U \begin{bmatrix} 1-(\frac{d^\beta}{\sigma_t})^2 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} U^T, \tag{B.11}$$

where U is an orthogonal matrix with the first column as u^β.

Thus, the Bias_a can be simplified as

$$\begin{aligned} \text{Bias}_a &= \frac{1}{\sigma_t^2} a'\tilde{\Gamma}(I_m - \frac{\beta\beta'}{\sigma_t^2})^{-1/2}\beta & \text{(B.12)}\\ &= \frac{d^\beta}{\sigma_t\sqrt{\sigma_t^2-(d^\beta)^2}} a'\tilde{\Gamma}u^\beta & \text{(B.13)}\\ &= \frac{1}{\sigma_t}\sqrt{\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}} a'\tilde{\Gamma}u^\beta, & \text{(B.14)}\\ &\leq \frac{1}{\sigma_t}\sqrt{\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}} \parallel a'\tilde{\Gamma} \parallel_2, & \text{(B.15)} \end{aligned}$$

where the bounds are reached when u^β is colinear with $a'\tilde{\Gamma}$, i.e., β is colinear with $a'\tilde{\Gamma}$.

Corollary 5.3.1 *Let d_1 be the largest singular value of $\tilde{\Gamma}$. For all $a \in \mathcal{R}^q$ with $\parallel a \parallel_2 = 1$,*

the confounding bias is bound by

$$Bias_a^2 \leq \frac{d_1^2}{\sigma_t^2} \frac{R_{T\sim U}^2}{1 - R_{T\sim U}^2}, \tag{B.16}$$

with equality when $a = u_1^{\tilde{\Gamma}}$*, the first left singular vector of* $\tilde{\Gamma}$*, and* β *being colinear with* $v_1^{\tilde{\Gamma}}$*, the first right singular vector of* $\tilde{\Gamma}$*. When* $a \in Null(\tilde{\Gamma})$*, the naive estimate is unbiased, that is,* $a'\tau^{naive} = a'\tau$*.*

Proof: From Equation B.14, we have

$$\text{Bias}_a = \frac{1}{\sigma_t} \sqrt{\frac{R_{T\sim U}^2}{1 - R_{T\sim U}^2}} a'\tilde{\Gamma}u^{\beta}, \tag{B.17}$$

where, according to Rayleigh quotient, $a'\tilde{\Gamma}u^{\beta}$ reaches its maximum, d_1, the largest singular value of $\tilde{\Gamma}$, when $a = u_1^{\tilde{\Gamma}}$, the first left singular vector of $\tilde{\Gamma}$, and $u^{\beta} = v_1^{\tilde{\Gamma}}$, the first right singular vector of $\tilde{\Gamma}$.

Thus,

$$\text{Bias}_a^2 \leq \frac{d_1^2}{\sigma_t^2} \frac{R_{T\sim U}^2}{1 - R_{T\sim U}^2}. \tag{B.18}$$

B.1.4 Proof of Proposition 5.4.1, Theorem 5.4.1 and Corollary

Proof of Proposition 5.4.1

Proposition 5.4.1 *Suppose there are c known null control outcomes* Y_j *with* $\tau_j = 0$ *for* $j \in \mathcal{C}$*. Then, the null control compatibility condition* $Q_{\tilde{\Gamma}_{\mathcal{C}}} \tau_j^{naive} = \tau_j^{naive}$ *must hold, where* $Q_{\tilde{\Gamma}_{\mathcal{C}}}$ *denotes the projection matrix into the column space of* $\tilde{\Gamma}_{\mathcal{C}}$*. In addition, the*

fraction of confounding variation in treatment is lower bounded by

$$R^2_{T\sim U} \geq R^2_{min} := \frac{\sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger} \tau_{\mathcal{C}}^{naive} \parallel_2^2}{1 + \sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger} \tau_{\mathcal{C}}^{naive} \parallel_2^2}, \tag{B.19}$$

where $\tilde{\Gamma}_{\mathcal{C}}^{\dagger}$ *denotes a generalized inverse of* $\tilde{\Gamma}_{\mathcal{C}}$.

Proof: Assume there are c null control outcomes, satisfying

$$\tau_{\mathcal{C}}^{\text{naive}} = \frac{1}{\sigma_t^2 \sqrt{1 - R^2_{T\sim U}}} \tilde{\Gamma}_{\mathcal{C}} \beta, \tag{B.20}$$

The solution for above equation exists if and only if $\tilde{\Gamma}_{\mathcal{C}} \tilde{\Gamma}_{\mathcal{C}}^{+} \tau_j^{\text{naive}} = \tau_j^{\text{naive}}$ holds, which ensures that null control assumptions are compatible. Under this condition, all solutions to Equation B.20 can be written as

$$\beta = \sigma_t^2 \sqrt{1 - R^2_{T\sim U}}\ \tilde{\Gamma}_{\mathcal{C}}^{+} \tau_{\mathcal{C}}^{\text{naive}} + (I - \tilde{\Gamma}_{\mathcal{C}}^{+} \tilde{\Gamma}_{\mathcal{C}}) w. \tag{B.21}$$

Since $\beta'\beta = \sigma_t^2 R^2_{T\sim U}$, w can be any $m \times 1$ vector satisfying

$$\parallel (I - \tilde{\Gamma}_{\mathcal{C}}^{+} \tilde{\Gamma}_{\mathcal{C}}) w \parallel_2^2 = \sigma_t^2 R^2_{T\sim U} - \sigma_t^4 (1 - R^2_{T\sim U}) \parallel \tilde{\Gamma}_{\mathcal{C}}^{+} \tau_{\mathcal{C}}^{\text{naive}} \parallel_2^2 \tag{B.22}$$

In addition,

$$\parallel (I - \tilde{\Gamma}_{\mathcal{C}}^{+} \tilde{\Gamma}_{\mathcal{C}}) w \parallel_2^2 \geq 0 \tag{B.23}$$

must hold, we know that $R^2_{T\sim U}$ must be at least

$$\frac{\sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger} \tau_{\mathcal{C}}^{\text{naive}} \parallel_2^2}{1 + \sigma_t^2 \parallel \tilde{\Gamma}_{\mathcal{C}}^{\dagger} \tau_{\mathcal{C}}^{\text{naive}} \parallel_2^2}. \tag{B.24}$$

Proof of Theorem 5.4.1 and Corollary

Theorem 5.4.1 *For any value of* $\mathbb{R}^2_{T\sim U} \geq R^2_{\min}$ *which satisfies null control compatibility condition, the confounding bias for the treatment effect of the interested outcome* $a'Y$ *is in the interval*

$$a'\tilde{\Gamma}\tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{naive}_{\mathcal{C}} \pm \sqrt{\frac{R^2_{T\sim U}}{\sigma_t^2(1-R^2_{T\sim U})} - \parallel \tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{naive}_{\mathcal{C}} \parallel_2^2} \parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} \parallel_2, \tag{B.25}$$

where $P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}}$ *is the* $m \times m$ *projection matrix into the complement of the row space of* $\tilde{\Gamma}_{\mathcal{C}}$.

Corollary *Under assumptions established in Theorem 5.4.1, null control outcomes reduce the width of the general worst-case ignorance region by a multiplicative factor of*

$$\sqrt{1 - \frac{R^2_{min}}{1-R^2_{min}} / \frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}} \frac{\parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} \parallel_2}{\parallel a'\tilde{\Gamma} \parallel_2} \leq 1. \tag{B.26}$$

Proof: For outcome of interest, $a'Y$, the omitted variable bias equals $\text{Bias}_a = \frac{1}{\sigma_t^2\sqrt{1-R^2_{T\sim U}}} a'\tilde{\Gamma}\beta$ and so it is bounded by

$$a'\tilde{\Gamma}\tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{\text{naive}}_{\mathcal{C}} \pm \sqrt{\frac{R^2_{T\sim U}}{\sigma_t^2(1-R^2_{T\sim U})} - \parallel \tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tau^{\text{naive}}_{\mathcal{C}} \parallel_2^2} \parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} \parallel_2 \tag{B.27}$$

with $P^{\perp}_{\tilde{\Gamma}_{\mathcal{C}}} := (I - \tilde{\Gamma}^{\dagger}_{\mathcal{C}}\tilde{\Gamma}_{\mathcal{C}})$, where the bounds are achieved when $(I - \tilde{\Gamma}^{+}_{\mathcal{C}}\tilde{\Gamma}_{\mathcal{C}})w$ has the largest cosine similarity with $a'\tilde{\Gamma}$.

Compare the bound with the one in Theorem 5.3.3, we see that the width of ignorance

region is shrunk by a multiplicative factor of

$$\sqrt{1-\frac{R^2_{\min}}{1-R^2_{\min}}/\frac{R^2_{T\sim U}}{1-R^2_{T\sim U}}}\frac{\parallel a'\tilde{\Gamma}P^{\perp}_{\tilde{\Gamma}_C}\parallel_2}{\parallel a'\tilde{\Gamma}\parallel_2}. \tag{B.28}$$

B.2 Additional Results

B.2.1 Additional Results from Analysis of Metabolomic Aging Clocks

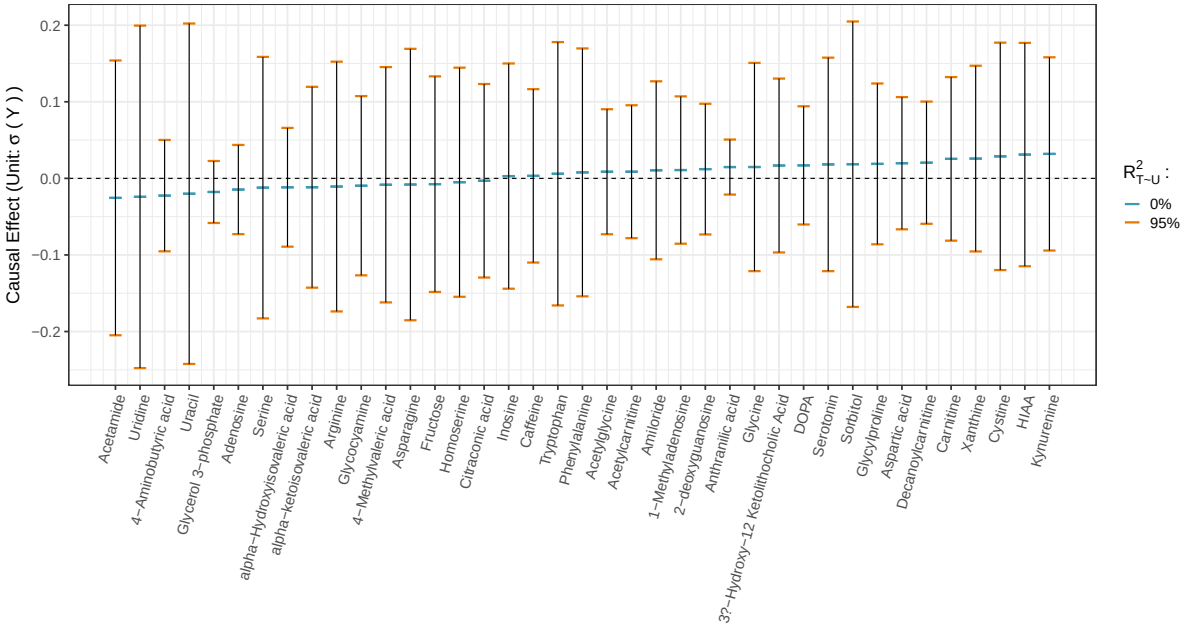


Figure B.1: Estimated effect of increasing age by one year on abundances of metabolites. The ignorance regions are on case by case basis with $R^2_{T\sim U} = 95\%$. All ignorance regions include zero.

Table B.1: Robustness Value for Metabolites

	Effect	RV(%)
Cystine	0.03	45
HIAA	0.03	49
Carnitine	0.03	56
Kynurenine	0.03	58
Xanthine	0.03	50
Decanoylcarnitine	0.02	62
DOPA	0.02	55
3?-Hydroxy-12 Ketolithocholic Acid	0.02	33
Serotonin	0.02	27
Aspartic acid	0.02	56
Glycylproline	0.02	43
Sorbitol	0.02	17
Amiloride	0.01	16
Acetylglycine	0.01	22
1-Methyladenosine	0.01	23
2-deoxyguanosine	0.01	33
Anthranilic acid	0.01	86
Acetylcarnitine	0.01	20
Glycine	0.01	21
Tryptophan	0.01	3
Phenylalanine	0.01	5
Inosine	0.00	1
Caffeine	0.00	2
Citraconic acid	0.00	1
Homoserine	-0.01	2
Glycocyamine	-0.01	13
Arginine	-0.01	9
4-Methylvaleric acid	-0.01	6
alpha-ketoisovaleric acid	-0.01	15
alpha-Hydroxyisovaleric acid	-0.01	36
Fructose	-0.01	6
Asparagine	-0.01	4
Adenosine	-0.01	64
Serine	-0.01	10
Glycerol 3-phosphate	-0.02	87
Uracil	-0.02	14
Uridine	-0.02	19
4-Aminobutyric acid	-0.02	71
Acetamide	-0.03	30

Bibliography

[1] J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder, *Smoking and lung cancer: recent evidence and a discussion of some questions*, *J. Nat. Cancer Inst* **22** (1959) 173–203.

[2] A. Franks, A. D'Amour, and A. Feller, *Flexible sensitivity analysis for observational studies without observable implications*, *Journal of the American Statistical Association* (2019), no. just-accepted 1–38.

[3] G. W. Imbens, *Sensitivity to exogeneity assumptions in program evaluation*, *American Economic Review* **93** (2003), no. 2 126–132.

[4] J. Pearl, *Causality*. Cambridge university press, 2009.

[5] J. Neyman, *On the application of probability theory to agricultural experiments. essay on principles. section 9*, *Statistical Science* **5** (1990 [1923]), no. 4 465–472.

[6] D. B. Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies.*, *Journal of educational Psychology* **66** (1974), no. 5 688.

[7] D. B. Rubin, *Comment*, *Journal of the American Statistical Association* **75** (1980), no. 371 591–593.

[8] P. R. Rosenbaum and D. B. Rubin, *Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome*, *Journal of the Royal Statistical Society. Series B (Methodological)* (1983) 212–218.

[9] T. J. VanderWeele and I. Shpitser, *On the definition of a confounder*, *Annals of statistics* **41** (2013), no. 1 196.

[10] M. Lechner, *Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption*, IZA Discussion Papers 91, Institute of Labor Economics (IZA), Dec., 1999.

[11] M. J. Lopez, R. Gutman, *et. al.*, *Estimation of causal effects with multiple treatments: a review and new ideas*, *Statistical Science* **32** (2017), no. 3 432–454.

[12] S. Greenland, *Basic methods for sensitivity analysis of biases*, *International journal of epidemiology* **25** (1996), no. 6 1107–1116.

[13] J. L. Gastwirth, A. M. Krieger, and P. R. Rosenbaum, *Dual and simultaneous sensitivity analysis for matched pairs*, *Biometrika* **85** (1998), no. 4 907–920.

[14] S. Vansteelandt, E. Goetghebeur, M. G. Kenward, and G. Molenberghs, *Ignorance and uncertainty regions as inferential tools in a sensitivity analysis*, *Statistica Sinica* **16** (2006), no. 3 953–979.

[15] T. J. VanderWeele and O. A. Arah, *Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders*, *Epidemiology* (2011) 42–52.

[16] T. J. VanderWeele, B. Mukherjee, and J. Chen, *Sensitivity analysis for interactions under unmeasured confounding*, *Statistics in medicine* **31** (2012), no. 22 2552–2564.

[17] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein, *Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models*, in *Statistical models in epidemiology, the environment, and clinical trials*, pp. 1–94. Springer, 2000.

[18] C. Cinelli and C. Hazlett, *Making sense of sensitivity: extending omitted variable bias*, *Journal of the Royal Statistical Society Series B (Statistical Methodology)* (12, 2019).

[19] V. Veitch and A. Zaveri, *Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding*, *arXiv preprint arXiv:2003.01747* (2020).

[20] V. Dorie, M. Harada, N. B. Carnegie, and J. Hill, *A flexible, interpretable framework for assessing sensitivity to unmeasured confounding*, *Statistics in Medicine* **35** (2016), no. 20 3453–3470, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6973].

[21] P. Gustafson, L. C. McCandless, *et. al.*, *When is a sensitivity parameter exactly that?*, *Statistical Science* **33** (2018), no. 1 86–95.

[22] J. M. Robins and R. D. Gill, *Non-response models for the analysis of non-monotone ignorable missing data*, *Statistics in medicine* **16** (1997), no. 1 39–56.

[23] M. J. Daniels and J. W. Hogan, *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.

[24] B. Zhang and E. J. T. Tchetgen, *A semiparametric approach to model-based sensitivity analysis in observational studies*, *arXiv preprint arXiv:1910.14130* (2019).

[25] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim, *Sensitivity analysis of linear structural causal models*, in *ICML*, 2019.

[26] Y. Lee and J. A. Nelder, *Hierarchical generalized linear models*, *Journal of the Royal Statistical Society: Series B (Methodological)* **58** (1996), no. 4 619–656.

[27] C. N. Morris and K. F. Lock, *Unifying the named natural exponential families and their relatives*, *The American Statistician* **63** (2009), no. 3 247–253, [https://doi.org/10.1198/tast.2009.08145].

[28] A. Franks, A. D'Amour, and A. Feller, *Flexible sensitivity analysis for observational studies without observable implications*, *arXiv e-prints* (Sep, 2018) arXiv:1809.00399, [arXiv:1809.0039].

[29] R. McCulloch, R. Sparapani, R. Gramacy, C. Spanbauer, and M. Pratola, *BART: Bayesian Additive Regression Trees*, 2018. R package version 1.9.

[30] A. M. Franks, E. M. Airoldi, and D. B. Rubin, *Non-standard conditionally specified models for non-ignorable missing data*, *arXiv preprint arXiv:1603.06045* (2016).

[31] A. R. Linero and M. J. Daniels, *Bayesian approaches for missing not at random outcome data: The role of identifying restrictions*, .

[32] M. Lechner, *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, in *Econometric Evaluation of Labour Market Policies* (M. Lechner and F. Pfeiffer, eds.), (Heidelberg), pp. 43–58, Physica-Verlag HD, 2001.

[33] H. A. Chipman, E. I. George, R. E. McCulloch, *et. al.*, *Bart: Bayesian additive regression trees*, *The Annals of Applied Statistics* **4** (2010), no. 1 266–298.

[34] R. D. McKelvey and W. Zavoina, *A statistical model for the analysis of ordinal level dependent variables*, *The Journal of Mathematical Sociology* **4** (1975), no. 1 103–120, [https://doi.org/10.1080/0022250X.1975.9989847].

[35] C. Gu, M. J. Lopez, and L. Hu, *The Estimation of Causal Effects of Multiple Treatments in Observational Studies Using Bayesian Additive Regression Trees*, *arXiv e-prints* (Jan, 2019) arXiv:1901.04312, [arXiv:1901.0431].

[36] Y. Wang and D. M. Blei, *The Blessings of Multiple Causes*, *arXiv e-prints* (May, 2018) arXiv:1805.06826, [arXiv:1805.0682].

[37] A. D'Amour, *Comment: Reflections on the deconfounder*, *Journal of the American Statistical Association* **114** (2019), no. 528 1597–1601.

[38] A. D'Amour, *On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative*, in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3478–3486, 2019.

[39] E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen, *Comment on "blessings of multiple causes"*, *Journal of the American Statistical Association* **114** (2019), no. 528 1611–1615.

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

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

[42] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, *Principal components analysis corrects for stratification in genome-wide association studies*, *Nature genetics* **38** (2006), no. 8 904–909.

[43] L. Zhang, Y. Wang, A. Ostropolets, J. J. Mulgrave, D. M. Blei, and G. Hripcsak, *The medical deconfounder: Assessing treatment effects with electronic health records*, *arXiv preprint arXiv:1904.02098* (2019).

[44] I. Bica, A. M. Alaa, C. Lambert, and M. van der Schaar, *From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges*, *Clinical Pharmacology & Therapeutics* (2020).

[45] W. Miao, W. Hu, E. L. Ogburn, and X. Zhou, *Identifying effects of multiple treatments in the presence of unmeasured confounding*, 2020.

[46] D. Kong, S. Yang, and L. Wang, *Multi-cause causal inference with unmeasured confounding and binary outcome*, *arXiv preprint arXiv:1907.13323* (2019).

[47] R. B. Nelsen, *An introduction to copulas*. Springer Science & Business Media, 2007.

[48] C. Cinelli, J. Ferwerda, and C. Hazlett, *sensemakr: Sensitivity analysis tools for ols in r and stata*, *Submitted to the Journal of Statistical Software* (2020).

[49] T. J. VanderWeele and P. Ding, *Sensitivity analysis in observational research: Introducing the e-value*, *Annals of Internal Medicine* **167** (July, 2017) 268.

[50] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, *Identifying causal effects with proxy variables of an unmeasured confounder*, *Biometrika* **105** (2018), no. 4 987–993.

[51] X. Shi, W. Miao, and E. T. Tchetgen, *A selective review of negative control methods in epidemiology*, *arXiv preprint arXiv:2009.05641* (2020).

[52] J. Yerushalmy, *The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations*, *International journal of epidemiology* **43** (2014), no. 5 1355–1366.

[53] E. Mitchell, R. Ford, A. Stewart, B. Taylor, D. Becroft, J. Thompson, R. Scragg, I. Hassall, D. Barry, E. Allen, *et. al.*, *Smoking and the sudden infant death syndrome*, *Pediatrics* **91** (1993), no. 5 893–896.

[54] L. Albers, C. Sobotzki, O. Kuß, T. Ajslev, R. F. Batista, H. Bettiol, B. Brabin, S. L. Buka, V. C. Cardoso, V. L. Clifton, *et. al.*, *Maternal smoking during pregnancy and offspring overweight: is there a dose–response relationship? an individual patient data meta-analysis*, *International Journal of Obesity* **42** (2018), no. 7 1249–1264.

[55] M.-J. A. Brion, S. D. Leary, G. D. Smith, and A. R. Ness, *Similar associations of parental prenatal smoking suggest child blood pressure is not influenced by intrauterine effects*, *Hypertension* **49** (2007), no. 6 1422–1428.

[56] G. D. Smith, *Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings?*, *Basic & clinical pharmacology & toxicology* **102** (2008), no. 2 245–256.

[57] B. K. Brew, T. Gong, D. M. Williams, H. Larsson, and C. Almqvist, *Using fathers as a negative control exposure to test the developmental origins of health and disease hypothesis: a case study on maternal distress and offspring asthma using swedish register data*, *Scandinavian journal of public health* **45** (2017), no. 17_suppl 36–40.

[58] A. E. Taylor, G. D. Smith, C. B. Bares, A. C. Edwards, and M. R. Munafò, *Partner smoking and maternal cotinine during pregnancy: implications for negative control methods*, *Drug and alcohol dependence* **139** (2014) 159–163.

[59] W. D. Flanders, M. Klein, L. A. Darrow, M. J. Strickland, S. E. Sarnat, J. A. Sarnat, L. A. Waller, A. Winquist, and P. E. Tolbert, *A method for detection of residual confounding in time-series and other observational studies*, *Epidemiology (Cambridge, Mass.)* **22** (2011), no. 1 59.

[60] W. D. Flanders, M. J. Strickland, and M. Klein, *A new method for partial correction of residual confounding in time-series and other observational studies*, *American journal of epidemiology* **185** (2017), no. 10 941–949.

[61] W. Miao and E. Tchetgen Tchetgen, *Invited commentary: bias attenuation and identification of causal effects with multiple negative controls*, *American journal of epidemiology* **185** (2017), no. 10 950–953.

[62] Y. Yu, H. Li, X. Sun, X. Liu, F. Yang, L. Hou, L. Liu, R. Yan, Y. Yu, M. Jing, *et. al.*, *Identification and estimation of causal effects using a negative-control exposure in time-series studies with applications to environmental epidemiology*, *American Journal of Epidemiology* **190** (2021), no. 3 468–476.

[63] M. Song, W. Hao, and J. D. Storey, *Testing for genetic associations in arbitrarily structured populations*, *Nature genetics* **47** (2015), no. 5 550–554.

[64] E. A. Boyle, Y. I. Li, and J. K. Pritchard, *An expanded view of complex traits: from polygenic to omnigenic*, *Cell* **169** (2017), no. 7 1177–1186.

[65] Kaggle, *Tmdb 5000 movie dataset*, Sep, 2017. data retrieved from Kaggle, `https://www.kaggle.com/tmdb/tmdb-movie-metadata`.

[66] J. Zheng, "Replication code for "copula-based sensitivity analysis for observational multi-treatment causal inference"." `https://github.com/JiajingZ/CopulaSensitivity`, 2021.

[67] J. Zheng, "Copsens: Copula-based sensitivity analysis method for unobserved confounding in multi-treatment inference.." `https://github.com/JiajingZ/CopSens`, 2021.

[68] J. Wang, Q. Zhao, T. Hastie, and A. B. Owen, *Confounder adjustment in multiple hypothesis testing*, *Annals of statistics* **45** (2017), no. 5 1863.

[69] L. J. O'Connor and A. L. Price, *Distinguishing genetic correlation from causation across 52 diseases and complex traits*, tech. rep., Nature Publishing Group, 2018.

[70] Q. Zhao, D. S. Small, and P. R. Rosenbaum, *Cross-screening in observational studies that test many hypotheses*, *Journal of the American Statistical Association* **113** (2018), no. 523 1070–1084.

[71] E. H. Kennedy, S. Kangovi, and N. Mitra, *Estimating scaled treatment effects with multiple outcomes*, *Statistical methods in medical research* **28** (2019), no. 4 1094–1104.

[72] T. J. VanderWeele, *Outcome-wide epidemiology*, *Epidemiology (Cambridge, Mass.)* **28** (2017), no. 3 399.

[73] B. Ning, S. Ghosal, and J. Thomas, *Bayesian method for causal inference in spatially-correlated multivariate time series*, *Bayesian Analysis* **14** (2019), no. 1 1–28.

[74] A. Mattei, F. Li, and F. Mealli, *Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program*, *The Annals of Applied Statistics* **7** (2013), no. 4 2336–2360.

[75] F. Mealli and B. Pacini, *Using secondary outcomes to sharpen inference in randomized experiments with noncompliance*, *Journal of the American Statistical Association* **108** (2013), no. 503 1120–1131.

[76] K. Steinhaeuser, N. Chawla, and A. Ganguly, *Improving inference of gaussian mixtures using auxiliary variables*, *Stat. Anal. Data Min* **8** (2015) 497–511.

[77] F. Mealli, B. Pacini, and E. Stanghellini, *Identification of principal causal effects using additional outcomes in concentration graphs*, *Journal of Educational and Behavioral Statistics* **41** (2016), no. 5 463–480.

[78] M. Lupparelli and A. Mattei, *Causal inference for binary non-independent outcomes*, *arXiv preprint arXiv:1710.07039* (2017).

[79] B. Jo and B. O. Muthén, *Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials*, in *New developments and techniques in structural equation modeling*, pp. 77–108. Psychology Press, 2001.

[80] M. A. Hernán, B. A. Brumback, and J. M. Robins, *Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures*, *Statistics in medicine* **21** (2002), no. 12 1689–1709.

[81] W. D. Flanders and M. Klein, *A general, multivariate definition of causal effects in epidemiology*, *Epidemiology* **26** (2015), no. 4 481–489.

[82] X. Li and P. Ding, *General forms of finite population central limit theorems with applications to causal inference*, *Journal of the American Statistical Association* **112** (2017), no. 520 1759–1769.

[83] N. Hwangbo, X. Zhang, D. Raftery, H. Gu, S.-C. Hu, T. J. Montine, J. F. Quinn, K. A. Chung, A. L. Hiller, D. Wang, *et. al.*, *An aging clock using metabolomic csf*, *bioRxiv* (2021).

[84] J. A. Gagnon-Bartsch and T. P. Speed, *Using control genes to correct for unwanted variation in microarray data*, *Biostatistics* **13** (2012), no. 3 539–552.

[85] J. A. Gagnon-Bartsch, L. Jacob, and T. P. Speed, *Removing unwanted variation from high dimensional data with negative controls*, *Berkeley: Tech Reports from Dep Stat Univ California* (2013) 1–112.

[86] Y. Behzadi, K. Restom, J. Liau, and T. T. Liu, *A component based noise correction method (compcor) for bold and perfusion based fmri*, *Neuroimage* **37** (2007), no. 1 90–101.

[87] J. Lucas, C. Carvalho, Q. Wang, A. Bild, J. R. Nevins, and M. West, *Sparse statistical modelling in gene expression genomics*, *Bayesian inference for gene expression and proteomics* **1** (2006), no. 1.

[88] Z. Wu and M. J. Aryee, *Subset quantile normalization using negative control features*, *Journal of Computational Biology* **17** (2010), no. 10 1385–1395.

[89] D. Schranner, G. Kastenmüller, M. Schönfelder, W. Römisch-Margl, and H. Wackerhage, *Metabolite concentration changes in humans after a bout of exercise: a systematic review of exercise metabolomics studies*, *Sports medicine-open* **6** (2020), no. 1 1–17.

[90] A. M. D. Livera, M. Sysi-Aho, L. Jacob, J. A. Gagnon-Bartsch, S. Castillo, J. A. Simpson, and T. P. Speed, *Statistical methods for handling unwanted variation in metabolomics data*, *Analytical chemistry* **87** (2015), no. 7 3606–3615.

[91] K. H. Gabbay, *The sorbitol pathway and the complications of diabetes*, *New England Journal of Medicine* **288** (1973), no. 16 831–836.

[92] A. Franks and J. Zheng, *Bayesian partial identification for muli-treatment inference with unobserved confounding*, ICML, 2021.

[93] X. Shi, W. Miao, J. C. Nelson, and E. J. Tchetgen Tchetgen, *Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** (2020), no. 2 521–540.

[94] W. Miao, X. Shi, and E. Tchetgen Tchetgen, *A confounding bridge approach for double negative control inference on causal effects*, *arXiv e-prints* (2018) arXiv–1808.

[95] R. Horn, *Matrix analysis*. Cambridge University Press, Cambridge Cambridgeshire New York, 1985.

[96] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic press, 1980.

[97] B. Everett, *An introduction to latent variable models*. Springer Science & Business Media, 2013.

[98] M. E. Tipping and C. M. Bishop, *Probabilistic principal component analysis*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** (1999), no. 3 611–622.

[99] T. P. Minka, *Automatic choice of dimensionality for pca*, in *Advances in neural information processing systems*, pp. 598–604, 2001.

[100] M. Gavish and D. L. Donoho, *The optimal hard threshold for singular values is* $4/\sqrt{3}$, *Information Theory, IEEE Transactions on* **60** (2014), no. 8 5040–5053.

[101] W. Hao, M. Song, and J. D. Storey, *Probabilistic models of genetic variation in structured populations applied to global human studies*, *Bioinformatics* **32** (2015), no. 5 713–721.

[102] P. Gopalan, J. M. Hofman, and D. M. Blei, *Scalable recommendation with poisson factorization*, *arXiv preprint arXiv:1311.1704* (2013).

[103] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, *Variational autoencoder for deep learning of images, labels and captions*, in *Advances in neural information processing systems*, pp. 2352–2360, 2016.

[104] R. Lopez, P. Boyeau, N. Yosef, M. I. Jordan, and J. Regier, *Decision-making with auto-encoding variational bayes*, *arXiv preprint arXiv:2002.07217* (2020).

[105] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, *Causal effect inference with deep latent-variable models*, in *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

[106] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black, and B. Scholkopf, *From variational to deterministic autoencoders*, in *International Conference on Learning Representations*, 2020.