Towards Reliable and Robust Causal Inference with High-dimensional Outcomes

Jin-Hong Du

July 2025

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Kathryn Roeder	(Chair)
Larry Wasserman	(MLD mentor)
Bryan Wilder	
Weijing Tang	
Jingshu Wang	(University of Chicago)

In fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics and Machine Learning.

To my family.

Abstract

Tens of thousands of simultaneous statistical hypothesis tests are routinely conducted in genomic studies to identify genes causally affected by disease. Recent advances in single-cell RNA sequencing and CRISPR technologies have enabled gene expression to be measured at high resolution. However, these data are often sparse, over-dispersed, and heterogeneous, posing substantial challenges for the reliable inference of multiple causal effects.

This thesis develops three complementary solutions.

(1) GCATE is a unified model-based framework for generalized linear models with latent confounding. By exploiting orthogonal structure and linear projections, GCATE enables consistent estimation and inference on direct effects under nonlinear models. In the high-dimensional regime when both sample and response sizes approach infinity, we derive Type-I error control of asymptotic z-tests and demonstrate false discovery rate control by the Benjamini-Hochberg procedure empirically. By comparing single-cell RNA-seq counts from two groups of samples, we demonstrate the suitability of adjusting confounding effects when significant covariates are absent from the model.

(2) causarray couples confounder estimation from GCATE and the semiparametric framework for multiple derived outcomes. The approach extends beyond average treatment effects to robust causal estimands and allows for flexible estimation using machine learning, and the resulting doubly robust pipeline maintains FDR or FDX control. Applications to an *in vivo* Perturb-seq screen of autism-risk genes and to three Alzheimer's transcriptomic datasets uncover clustered neuronal pathways implicated in disease.

(3) Pii supplies assumption-lean post-integration inference by leveraging negative control outcomes to adjust latent heterogeneity. The resulting doubly robust estimators achieve consistency and efficiency under weak conditions, enabling inference after integration with machine learning for data-adaptive estimation. The empirical performance is evaluated via simulations using random forests and further demonstrated on single-cell CRISPR datasets with potential unmeasured confounding.

Together, these methods form a principled toolkit for causal inference in complex genomic settings, addressing non-Gaussianity, heterogeneity, high-dimensionality, and unmeasured confounding, and enabling reliable discovery of disease-related genes and pathways.

Acknowledgments

This thesis marks the end of a long journey of discovery, learning, and growth—a journey made possible only through the guidance, generosity, and friendship of many remarkable people. I am profoundly grateful to each of them.

Advisors. First and foremost, I would like to express my profound gratitude to my advisors, Kathryn Roeder and Larry Wasserman. Their unwavering support, incisive feedback, and high scholarly standards have shaped every page of this dissertation. I have been uniquely fortunate to learn from mentors whose expertise spans the entire spectrum of our discipline: Kathryn, a leading authority in statistical genetics, and Larry, who knows all of statistics, all of nonparametric statistics, and all of causation. Their complementary perspectives—application and theory—have defined how I think about research in statistics and will continue to guide me well beyond this thesis.

Thesis committee. I am indebted to Bryan Wilder, Weijing Tang, and Jingshu Wang for serving on my committee. Their incisive questions, detailed feedback, and thoughtful suggestions improved both the depth and the presentation of this work; many of their comments are reflected in the pages that follow.

Collaborators. Research thrives on collaboration, and I have been fortunate to work with outstanding colleagues. (1) Statistical methods and applications for single-cell data analysis: Zhanrui Cai, Edward H. Kennedy, Jing Lei, Hansruedi Mathys, Haeun Moon, Maya Shen, Zhenghao Zeng, Yaoming Zhen, Wenbin Zhou. (2) Over-parameterized learning theory: Pierre C. Bellec, Arun Kumar Kuchibhotla, Takuya Koriyama, Pratik Patil, Alessandro Rinaldo, Kai Tan, Ryan J. Tibshirani. (3) Earlier projects before my doctorial studies: portfolio selection with Yifeng Guo and Xueqin Wang; multiparameter eigenvalue problems with José Israel Rodríguez and Lek-Heng Lim; and single-cell trajectory inference with Tianyu Chen and Jingshu Wang. I thank each of you for sharing your insight, patience, and excitement for research.

Community and friends. Life at CMU was immeasurably richer thanks to a supportive cohort and a vibrant scholarly community. Special thanks to Bernie Devlin, Lambertus Klei, Catherine Wang, F. William Townes, Tianyu Chen, and all other members of the CMU GenStats Lab Group; and to Zach Branson, Eli Ben Michael and all other members of the CMU Causal Reading Group for suggestions and many insightful conversations. I am deeply grateful for the Statistics department's robust computing resources, especially our high-performance servers, and for the prompt, expert assistance that Jake Gordon and Carl Skipper unfailingly provided whenever technical issues arose. Additionally, I am grateful to the Machine Learning Department for fostering a vibrant community through its retreats, parties, and other engaging social events. As an international student, I really appreciated the group of friends from my home country that made Pittsburgh feel like home. Finally, my undergraduate roommates—Xiuwen Duan, Tao Dong, Xingyu Fu, and Yifeng Guo—have been unfailing sources of humour, perspective, and encouragement since our first year together; their friendship has sustained me through both undergraduate and graduate life and will remain a lifelong treasure.

Family. Above all, I owe my deepest gratitude to my family. To my parents and my sister: your unconditional love, patience, and belief in me have been my foundation and my strength. This dissertation would not exist without your steadfast support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of genomic research has transformed our understanding of biological processes and disease mechanisms. Advances in single-cell RNA sequencing (scRNA-seq) have driven this rapid progress, offering unprecedented insights into gene expression patterns at the cellular level [161]. The high resolution provided by scRNA-seq data is essential to elucidate cellular heterogeneity and its implications for health and disease [50, 110, 164]. However, fully harnessing the potential of these data requires robust analytical frameworks capable of moving beyond association to unravel complex causal relationships at single-cell resolution [49, 92, 134]. The fundamental difference between association and causation is that association assesses correlations between treatments and outcomes, whereas causal inference aims to quantify the effect of a treatment on an outcome. A popular framework for causal inference is the *potential outcomes* framework, which estimates what would have happened if a different treatment had been assigned, the *counterfactual* [49, 71]. Causal inferences are crucial for understanding biological processes and disease mechanisms, with important implications for treatments, precision medicine, genomic medicine, and related fields [149, 155].

One of the primary challenges in leveraging scRNA-seq data for causal inference is its inherent hierarchical organization and heterogeneity [38, 49, 134]. Cells from the same individual are not independent observations. They share biological factors, such as correlated gene expression, and technical factors, including batch effects introduced during storage and sequencing. These dependencies violate the assumption of independent and identically distributed (i.i.d.) samples, complicating statistical analyses and rendering traditional methods inadequate for handling heterogeneous data with unwanted variations [140, 147]. Furthermore, most genomic studies are observational in nature. Unlike randomized controlled trials, observational studies lack complete knowledge of the disease or treatment assignment mechanism, leading to potential biases in counterfactual estimation.

CRISPR perturbation experiments, a more recent but rapidly expanding area, offer a new set of challenging analysis scenarios [29, 68, 82]. For this experimental setting, perturbed cells are contrasted with cells that receive a non-targeting perturbation. While there is some randomness in the treatment assignment, it is not entirely random: continuous unmeasured confounders such as variability in cell size or differential drug exposure can result in biased causal estimates. Additionally, when such experiments are performed in vivo, the possibility of confounding increases [79], further justifying the need for robust causal inference analysis.

In summary, the challenges of reliable causal inference for scRNA-seq and CRISPR analysis lie in:

(1) **Noisy**: scRNA-seq data exhibit *sparsity*, with a significant portion of the data consists of zeros, indicating no detected expression for many genes across cells; and *overdispersion*, with the variance of gene expression counts surpassing the mean, challenging conventional assumptions of Poisson or binomial distributions for count data modeling.

(2) **High-dimensional and intricate correlated**: scRNA-seq data are inherently high-dimensional, with the number of measured genes (variables) far exceeding the number of cells (observations), and this high-dimensionality is coupled with correlations among gene expressions that require sophisticated joint modeling approaches for consistent estimation and valid inference.

(3) **Observational**: As these studies are typically observational, scRNA-seq data come with inherent potential confounders and heterogeneity among cells or subjects. This heterogeneity necessitates the deployment of advanced causal inference tools to manage these challenges effectively, ensuring accurate interpretation and conclusions from both bulk-cell and single-cell data analyses.

These characteristics necessitate the development and application of advanced statistical causal inference methods that can jointly model these aspects, addressing sparsity, overdispersion, high-dimensionality, correlation, and heterogeneity to extract meaningful biological insights from single-cell data.

The problems of modelling and predicting the single-cell gene expressions have been extensively studied in various works. For instance, Du et al. [41, 42] consider estimation and model selection with ensemble methods, with the theoretical understandings of overparameterized ensemble learning explored in a series of work [13, 135, 136], and Zhen and Du [187] consider neighborhood prediction model using network information. The probabilistic deep generative modeling of multi-modal single-cell datasets is also explored by Du et al. [40, 43], Moon et al. [128], Zhou and Du [188]. A natural next step is to establish valid statistical inference for identifying differentially expressed and causally expressed genes. In genomics, only a few methods have been proposed for drawing conclusions about causal gene identification. The lack of reliable and robust causal inference approaches for genomics discoveries inspires us to study the inferential methods for causal inference with multiple outcomes in genomics.

In response to these challenges, this thesis explores three approaches for simultaneous causal inference on multiple genes:

(1) GCATE [48]: a model-based framework for multivariate generalized linear models with latent confounding. GCATE projects out unmeasured factors, applies sparsity-aware bias correction, and provides valid large-scale z–tests with false-discovery-rate (FDR) control even for high-dimensional, over-dispersed count data.

(2) causarray [47, 49]: a two-stage procedure that (i) uses GCATE to estimate unobserved confounders and (ii) combines these estimates with doubly robust semiparametric estimators to obtain causal contrasts. The pipeline accommodates flexible machine-learning nuisances, handles heterogeneous single-cell and pseudo-bulk outcomes, and supports both FDR and family-wise discovery-rate (FDX) control.

(3) Pii [45]: an assumption-lean *post-integration inference* approach that adjusts for latent heterogeneity via negative control outcomes. Pii delivers deterministic bias corrections and doubly robust estimators that remain consistent and efficient under model misspecification, mediation, and moderation, enabling reliable inference after data-adaptive integration.

The three methods will be detailed in the following three chapters, presented in order.

Chapter 2

Simultaneous inference for generalized linear models with unmeasured confounders

Material in this chapter first appeared as Du et al. [48].

2.1 Introduction

To discover genes that are differentially expressed under different experimental conditions or across groups of samples, large numbers of simultaneous hypothesis tests must be performed. These tests are made more challenging by the presence of unmeasured covariates that bias the analyses. In 2007, Leek and Storey [96, 97] presented their pathbreaking "surrogate variable" approach to control for unmeasured confounding effects in differential expression (DE) studies using microarray data. These confounders go by various names in the literature, including batch effects, surrogate variables, latent effects, or simply unwanted variations [56, 98, 160]. Adjusting for confounding effects is crucial because they may distort the correct null distribution of the test statistics, and consequently, standard statistical approaches can be substantially biased [119, 175]. Due to burgeoning developments in the genomics field, DE testing has been dramatically expanded to include a variety of genomic readouts beyond microarray, in which the normality of the observed counts rarely holds. Inspired by modern-day omic studies, the concerns about confounding are more urgent than ever, and there is a pressing need to adapt statistical approaches to changing data types.

The problem of confounder adjustment has been an important topic in statistics in recent years. To characterize the confounding effects, the pioneering work in this field assumes a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top + \boldsymbol{E}$, where $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ is the gene expression matrix, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is the measured covariate matrix, $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ is the direct effect to be estimated, $\boldsymbol{Z} \in \mathbb{R}^{n\times r}$ is the latent factor matrix, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$ is the latent factor loading, and $\boldsymbol{E} \in \mathbb{R}^{n\times p}$ is the additive noise. The early investigations study the statistical inference problem under this model by further imposing a linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Z}$, assuming either $\boldsymbol{X}$ causes $\boldsymbol{Z}$ as in Figure 2.1(a), i.e., $\boldsymbol{Z}$ is a hidden mediator [59, 97, 175], or $\boldsymbol{Z}$ causes $\boldsymbol{X}$ as in Figure 2.1(b), i.e., $\boldsymbol{Z}$ is a hidden confounder [62, 159].

In the presence of hidden mediators, where the observed covariates are the cause of the hidden variables, Wang et al. [175] and Gerard and Stephens [59] study the statistical inference problem

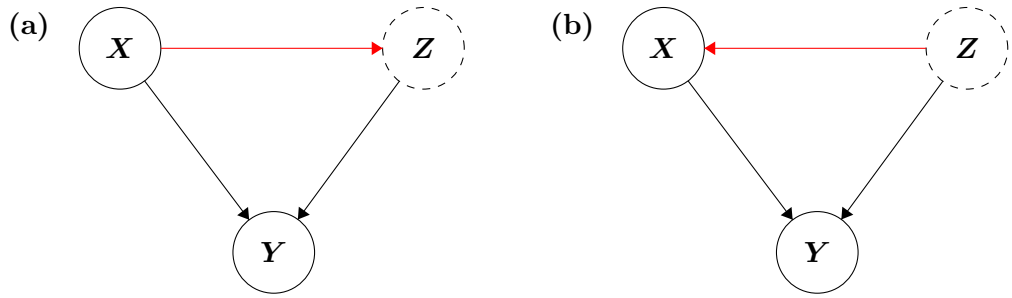

Figure 2.1: Causal diagrams on the generative models illustrating the relationship between the covariate $\boldsymbol{X}$, the latent variable $\boldsymbol{Z}$, and the response $\boldsymbol{Y}$. **(a)** $\boldsymbol{Z}$ is a hidden mediator when $\boldsymbol{X}$ causes $\boldsymbol{Z}$. **(b)** hidden confounder when $\boldsymbol{Z}$ causes $\boldsymbol{X}$. Note that we do not require knowledge of the relationship between $\boldsymbol{X}$ and $\boldsymbol{Z}$ for the analysis in this paper.

for multiple outcomes ($p > 1$) by assuming a linear relationship between the observed variables and the hidden variables. In the context of hidden confounders and a single outcome ($p = 1$), Guo et al. [62] propose a doubly debiased lasso estimator and establish asymptotic normality; Ćevid et al. [22] propose a spectral de-confounding method; and Sun et al. [159] analyze non-asymptotic and asymptotic false discovery control with high-dimensional covariates.

More recently, methods for estimating primary effects extend beyond linear dependence structures between covariates and confounders. For instance, Jiang and Ning [76] model the interaction between the covariates and the confounders, and projection-based methods are employed to estimate the primary effects under arbitrary dependency [19, 95, 120]. For statistical inference, McKennan and Nicolae [119] propose an estimator that is asymptotically equivalent to the ordinary least squares estimators obtained when every covariate is observed, and Bing et al. [20] establish asymptotic normality, efficiency, and consistency.

The applicability of the aforementioned methods to the nonlinear model remains challenging. Limited research has been done to address adjustments for confounding effects under the setting of *arbitrary confounding mechanisms*, *nonlinear models*, and *multiple outcomes*. For empirical studies, Salim et al. [147] propose a heuristic algorithm that utilizes a pseudo-replicate design matrix and negative control genes to remove unwanted variations. For theoretical analysis, to the best of our knowledge, the related literature that explores slightly broader settings is limited to Feng [51], who studies nonlinear factor models concerning treatment effects with a single outcome by PCA-based matching, and Ouyang et al. [131], who study the generalized linear models with a single outcome and linear hidden confounders. However, both of these works assume the covariates are some functions of the unobserved confounders.

Our work is inspired by the rapid developments in the field of genomics, particularly single-cell omics [50]. For example, CRISPR perturbations with single-cell sequencing readouts have promised extraordinary scientific insight [29, 68, 82]; due to the sparsity of outcomes and the nature of the molecular readout, these data are not suitable for analysis by linear models under Gaussianity assumptions [12, 150]. Hence, our development of generalized linear models for confounding is timely.

In this paper, we adopt the term "confounder" to encompass a broad category of latent variables, including both mediators and confounders, as defined in the context of causal inference literature. The purpose of this paper is to derive valid simultaneous inference for multivariate generalized linear models in the presence of unmeasured confounding effects. Existing methods in

this domain typically focus on Gaussian linear models [19, 20, 175] or necessitate direct modeling of the relationship between covariates and confounders [51, 131]. To the best of our knowledge, the proposed method is the first estimation and inference framework capable of (1) accommodating general relationships between observed covariates and unmeasured confounders, allowing for *arbitrary confounding mechanisms*; (2) utilizing generalized linear models, allowing for *nonlinear modeling*; and (3) incorporating information from *multiple outcomes*. Our approach leverages the orthogonal structures inherent in the problem, incorporating linear projection techniques into both estimation and inference processes to effectively mitigate confounding effects and elucidate primary effects. Notably, it exhibits significant utility in high-dimensional sparse count data, as demonstrated through the analysis of single-cell datasets on systemic lupus erythematosus disease in Section 2.6.

Our proposed procedure GCATE (generalized confounder adjustment for testing and estimation) consists of three main steps. In the first step, we use joint maximum likelihood estimation [26, 27] to obtain the initial estimate of the marginal effects and uncorrelated latent components by projecting the latent factor $\boldsymbol{Z}$ to the orthogonal space of $\boldsymbol{X}$, from which we recover the column space of $\boldsymbol{\Gamma}$. In the second step, we use a similar strategy to obtain the estimates of both $\boldsymbol{Z}$ and primary effect $\boldsymbol{B}$, by constraining the latter to be orthogonal to the estimated latent coefficients $\widehat{\boldsymbol{\Gamma}}$ and using ℓ_1-regularization to encourage sparsity. Lastly, the valid inference is guaranteed by a bias-corrected estimator of $\widehat{\boldsymbol{B}}$, which innovates a link-specific weight function, similar to [21, 74], while incorporating projection-based score adjustments that combine the information from multivariate responses.

In our theoretical framework of confounded generalized linear models, we establish conditions for identifying the latent coefficients and direct effects. Furthermore, we provide non-asymptotic estimation error bounds for these estimated quantities in high-dimensional scenarios where both the sample size n and response size p tend to be infinity. In particular, we derive element-wise ℓ_2-norm and ℓ_1-norm bounds for the estimation error of the primary effects by effectively controlling the column-wise estimation errors of the latent components. Lastly, we demonstrate the asymptotic normality of our proposed bias-corrected estimator and show the proper control of statistical errors, thereby enabling the construction of valid confidence intervals and hypothesis tests.

Organization and Notation. In Section 2.2, we set up our modeling framework, which extends existing results in the literature to the generalized linear model setting. In Section 2.3, we describe our strategy for estimation and establish bounds on the estimation error of the parameters of interest. In Section 2.4, we motivate and construct asymptotically valid confidence intervals and hypothesis tests. Finally, in Section 2.5 and Section 2.6, we study the empirical behavior of our estimators in realistic simulations and a study of gene expression in lupus patients. Technical proof of the results is provided in the supplementary material.

Throughout our exposition, we will use the following notational conventions. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, we use $\boldsymbol{a}_i$, $\boldsymbol{A}_j$, and a_{ij} to denote its ith row, jth column, and (i,j)-th entry, respectively, for $i = 1,\ldots,n$, $j = 1,\ldots,p$. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ with full column rank, let $\mathcal{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$ and $\mathcal{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I}_p - \mathcal{P}_{\boldsymbol{A}}$ be the orthogonal projection matrices on the $\boldsymbol{A}$'s column space and its orthogonal space, respectively. For any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\lambda_i(\boldsymbol{A})$ denotes its ith largest eigenvalue. The symbol "$\odot$" denotes the Hadamard product. We use "o" and "$\mathcal{O}$" to denote the little-o and big-O notations and let "$o_{\mathbb{P}}$" and "$\mathcal{O}_{\mathbb{P}}$" be their probabilistic counterparts. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ or $b_n \gg a_n$ if $a_n = o(b_n)$; $a_n \lesssim b_n$

or $b_n \gtrsim a_n$ if $a_n = \mathcal{O}(b_n)$; and $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Convergence in distribution and probability are denoted by "$\xrightarrow{\text{d}}$" and "$\xrightarrow{\text{p}}$".

2.2 Modeling Differential Expression

In the context of DE testing and related applied problems, the outcome variable can be a variety of measures, including gene expression, protein abundance, and open chromatin. For simplicity of exposition, we will describe our methods in the context of tests for differential gene expression. These tests aim to contrast outcomes from case versus control samples, wherein case and control observations may be derived from various study designs, spanning the spectrum from diseased versus healthy subjects to perturbed versus non-targeted cells.

2.2.1 Generalized linear model with hidden confounders

Suppose the gene expression $\boldsymbol{y} \in \mathbb{R}^p$ is a p-dimensional random vector containing conditional independent entries from a one-dimensional exponential family with density:

$$p(y_j \mid \theta_j) = h(y_j)\ \exp\left(y_j\theta_j - A(\theta_j)\right),$$

where $\theta_j \in \mathbb{R}$ is the *natural parameter*, and $A(\cdot)$ and $h(\cdot)$ are functions that depend on the member of the exponential family. We restrict ourselves to the regular families whose natural parameter space is a nonempty open set and A is continuously thrice differentiable, which is satisfied by most common exponential families as summarized in Table A.61. Because the one-parameter exponential family is minimal, the natural parameter space is convex, and the *log-partition function* A is strictly convex. If we know the distribution of $\boldsymbol{y}$, then $\boldsymbol{\theta} \in \mathbb{R}^p$ is a unique solution to the equation $\mathbb{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}] = A'(\boldsymbol{\theta})$, where A' is the first derivative of A and applied element-wise to $\boldsymbol{\theta}$; equivalently, $\boldsymbol{\theta} = A'^{-1}(\mathbb{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}])$. In other words, we can recover $\boldsymbol{\theta}$ based on the information of the first moment of $\boldsymbol{y}$ and the log-partition function A.

To associate multiple outcomes with both covariates and hidden confounders, one can naturally consider the generalized linear model, where the natural parameters are linear functions of both the observed covariates $\boldsymbol{x} \in \mathbb{R}^d$ and the unmeasured confounder $\boldsymbol{z} \in \mathbb{R}^r$:

$$\boldsymbol{\theta}_{p\times 1} = \boldsymbol{B}_{p\times d}\boldsymbol{x}_{d\times 1} + \boldsymbol{\Gamma}_{p\times r}\boldsymbol{z}_{r\times 1}.$$

Here, $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ are the linear coefficients. Denote $\boldsymbol{D}\boldsymbol{x}$ the linear projection of $\boldsymbol{z}$ onto $\boldsymbol{x}$, where $\boldsymbol{D} := \mathbb{E}[\boldsymbol{z}\boldsymbol{x}^\top]\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1} \in \mathbb{R}^{r\times d}$ is the projection coefficient and $\boldsymbol{w} = \boldsymbol{z} - \boldsymbol{D}\boldsymbol{x}$ is the residual uncorrelated with $\boldsymbol{x}$. To see how $\boldsymbol{z}$ may affect the inference on $\boldsymbol{B}$, note that

$$\boldsymbol{\theta} = (\boldsymbol{B} + \boldsymbol{\Gamma}\boldsymbol{D})\boldsymbol{x} + \boldsymbol{\Gamma}\boldsymbol{w}. \tag{2.2.1}$$

When $\boldsymbol{y}$ is normally distributed, the confounding effects occur even when regressing the mean response $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{y} \mid \boldsymbol{\theta}]$ on $\boldsymbol{x}$, which yields the *confounded coefficient* $\boldsymbol{B} + \boldsymbol{\Gamma}\boldsymbol{D}$ while the *direct effect* of interest is $\boldsymbol{B}$. When $\boldsymbol{y}$ comes from general exponential families, the confounding effects are more intractable because all moments and cumulants of the response may be affected by the colinearity of $\boldsymbol{x}$ and $\boldsymbol{z}$.

In the context of genomic analysis, the problem of confounding is more severe when the number of available covariates is limited. In particular, one typically encounters high-dimensional scenarios characterized by a substantial number of genes, often surpassing the available numbers

of covariates and hidden confounding factors. In this paper, we also consider such a challenging scenario where the number of genes is much larger than the numbers of the observed covariates and the unmeasured confounders, namely, $p \gg d$ and $p \gg r$. Under such challenges, the first natural and essential question one may ask is whether there is any hope to disentangle the confounding effects and identify the direct effects.

The answer to this inquiry is affirmative. On the one hand, the column space of $\boldsymbol{\Gamma}$ can be identified up to rotations if $\mathrm{Cov}(\boldsymbol{\Gamma}\boldsymbol{w}) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ has rank r, where $\boldsymbol{\Sigma}_w = \mathrm{Cov}(\boldsymbol{w})$ is the covariance of the uncorrelated latent factors. This fact originates from basic principles in linear algebra, frequently employed in factor analysis [7, 8]. On the other hand, once the column space of $\boldsymbol{\Gamma}$ is known, one can apply the orthogonal projections to remove the confounding effects based on (2.2.1):

$$\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{\theta} = \mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{B}\boldsymbol{x},$$

where $\mathcal{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top$ and $\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp} = \boldsymbol{I}_p - \mathcal{P}_{\boldsymbol{\Gamma}}$ are the orthogonal projection matrices that project vectors on to the image of $\boldsymbol{\Gamma}$ and the orthogonal complement of $\boldsymbol{\Gamma}$, respectively. By regressing $\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{\theta}$ on $\boldsymbol{x}$, we can further obtain the unconfounded primary effects $\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{B} = \mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\mathbb{E}[\boldsymbol{\theta}\boldsymbol{x}^\top]\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}$. However, the other component, $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}$, often poses challenges in identifiability unless additional conditions are imposed. Typically, extra assumptions on the spectrum of $\boldsymbol{\Gamma}$ and the sparsity of $\boldsymbol{B}$ are necessary, to assert that $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ are asymptotically orthogonal, in the sense that $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}$ is negligible [20, 95, 175]. In that case, $\boldsymbol{B}$ can be well approximated by $\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{B}$. Below, we give one sufficient identification condition.

Proposition 1 (Identification of $\boldsymbol{B}$)**.** Suppose there exists a sequence $\{\tau_p\}_{p\in\mathbb{N}}$ that is uniformly lower bounded away from zero such that the following conditions hold:

$$\lambda_r(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top) \geq \tau_p, \qquad \max_{1\leq j\leq p}(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top)_{jj} = \mathcal{O}(1), \qquad \max_{1\leq \ell\leq d}\|\boldsymbol{B}_\ell\|_1 = o(\tau_p). \tag{2.2.2}$$

Then as p tends to infinity, it follows that $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{B} + \boldsymbol{o}(1)$ and $\|\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}\|_{\mathrm{F}} \lesssim \sqrt{p}\|\boldsymbol{B}\|_{1,1}/\tau_p$, where $\|\cdot\|_{1,1}$ is the element-wise ℓ_1-norm. Further, $\mathcal{P}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{B}$ can be identified from the first two moments of $\boldsymbol{x}, \boldsymbol{y}$ asymptotically.

As hinted above, the lower bound condition of $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$'s spectrum in (2.2.2) ensures that the column space of $\boldsymbol{\Gamma}$ can be identified up to rotations. The second condition guarantees that the diagonal entries of $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ are balanced. Finally, the last condition in (2.2.2) can hold when $\boldsymbol{B}$ is sparse, and its entry is bounded. Compared to Bing et al. [20, Theorem 1] where the response $\boldsymbol{y}$ is normally distributed, the identifiability condition of Proposition 1 applies for exponential families, which is of much broader generality. Furthermore, the smallest eigenvalue of $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ can grow at a specific rate τ_p in Proposition 1. When $\tau_p = p$, we can recover the result in Bing et al. [20, Theorem 1]. Lastly, we also provide a norm bound for the residual $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}$, which is helpful for later analysis of the estimation errors.

2.2.2 Random samples

While the preceding identification results are applicable when population moments are known, practical scenarios involve the observation of independent and identically distributed (i.i.d.) samples. Consequently, statistical estimation becomes imperative to disentangle direct effects from the confounding effects based on samples. Suppose $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \ldots, n$ are n i.i.d. samples coming from the same distribution as $(\boldsymbol{x}, \boldsymbol{y})$, and let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ denote the design

matrix and the gene expression matrix, respectively. The expression y_{ij} of the ith observation and the jth gene has the density:

$$p(y_{ij} \mid \theta_{ij}) = h(y_{ij}) \exp\left(y_{ij}\theta_{ij} - A(\theta_{ij})\right),$$

where θ_{ij} is the natural parameter. In matrix form, the natural parameters decompose as

$$\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top,$$

where $\boldsymbol{B} \in \mathbb{R}^{p\times d}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times r}$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$ are unknown. Note that y_{ij}'s are conditionally independent given the natural parameter $\boldsymbol{\Theta}$.

One natural way to estimate the unknown variable $\boldsymbol{Z}$ and parameters $(\boldsymbol{B}, \boldsymbol{\Gamma})$ is to perform maximum likelihood estimation. Ignoring the constant terms, the negative log-likelihood function of $\boldsymbol{Y}$ is given by

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{\Gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}\theta_{ij} - A(\theta_{ij})). \tag{2.2.3}$$

The second notation $\mathcal{L}(\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{\Gamma})$ reflects the dependence of $\boldsymbol{\Theta}$ on the model parameters $\boldsymbol{B}$ and unknown quantities $\boldsymbol{\Gamma}, \boldsymbol{Z}$. Addressing the challenges of nonconvexity and high dimensionality requires developing efficient algorithms to estimate these unknown quantities and analyze their statistical properties.

To overcome the difficulty of estimation, one critical observation comes from the projection-based decomposition and Proposition 1:

$$\begin{aligned}
\boldsymbol{\Theta} &= \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top \\
&= (\boldsymbol{X}\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp} + \boldsymbol{X}\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}) + (\boldsymbol{X}\boldsymbol{D}^\top\boldsymbol{\Gamma}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top) \\
&= \boldsymbol{X}\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp} + \mathcal{P}_{\boldsymbol{X}}\boldsymbol{Z}\boldsymbol{\Gamma}^\top + \mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{W}\boldsymbol{\Gamma}^\top + \boldsymbol{o}_{\mathbb{P}}(1),
\end{aligned}$$

where we replace the best linear projection $\boldsymbol{X}\boldsymbol{D}$ with its empirical counterpart $\mathcal{P}_{\boldsymbol{X}}\boldsymbol{Z}$ in finite samples, which yield negligible terms that contribute to $\boldsymbol{o}_{\mathbb{P}}(1)$. It is worth noting that $\boldsymbol{X}\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}$ and $\mathcal{P}_{\boldsymbol{X}}\boldsymbol{Z}\boldsymbol{\Gamma}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top$ have orthogonal columns, while $\boldsymbol{X}\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp} + \mathcal{P}_{\boldsymbol{X}}\boldsymbol{Z}\boldsymbol{\Gamma}^\top$ and $\mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{W}\boldsymbol{\Gamma}^\top$ have orthogonal rows. Our analysis will then take advantage of such two-way structural orthogonality to perform both estimation and inference for the parameters of interest, as detailed in the following sections.

2.3 Estimation

From now on, we will use an asterisk on the upper subscript to indicate the population parameters and the true latent factors. Specifically, we denote the underlying parameter as

$$\boldsymbol{\Theta}^* = \boldsymbol{X}\boldsymbol{B}^{*\top} + \boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top} = \boldsymbol{X}(\boldsymbol{B}^* + \boldsymbol{\Gamma}^*\boldsymbol{D}^*)^\top + \boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}.$$

Let $\mathcal{R} \subseteq \mathbb{R}$ be an open domain of θ such that $A(\theta) < \infty$ for all $\theta \in \mathcal{R}$. For a given $C > 0$, define $\mathcal{R}_C = \mathcal{R}\cap[-C, C]$ for Gaussian, Binomial and Poisson distributions and $\mathcal{R}_C = \mathcal{R}\cap[-C, -1/C]$ for Negative Binomial distributions. For our theoretical results, we assume the existence of constant $C > 1$ such that the following common assumptions hold.

Assumption 1 (Model parameters)**.** Assume that $\boldsymbol{\Theta}^* \in \mathcal{R}_C^{n\times p}$ with probability ι_n for some deterministic sequence ι_n tending to one as n tends to infinity. The primary coefficient satisfies that $\max_{1\le \ell\le d} \|\boldsymbol{B}_\ell^*\|_0 \le s$ for some $1 \le s \le p$ and $\max_{1\le j\le p} \|\boldsymbol{b}_j^*\|_2 \le C$.

Assumption 2 (Covariates)**.** Assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. ν-sub-Gaussian random vectors with second moment $\boldsymbol{\Sigma}_x := \mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top]$ such that $C^{-1} \le \lambda_p(\boldsymbol{\Sigma}_x) \le \lambda_1(\boldsymbol{\Sigma}_x) \le C$.

Assumption 3 (Latent vectors)**.** Assume that the uncorrelated latent factors $\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_n^*$ are i.i.d. ν-sub-Gaussian random vectors with zero means and covariance $\boldsymbol{\Sigma}_w$, such that $C^{-1} \le \lambda_r(\boldsymbol{\Sigma}_w) \le \lambda_1(\boldsymbol{\Sigma}_w) \le C$; and the factor loadings $\boldsymbol{\Gamma}^*$ satisfy that $C^{-1} \le \lambda_r(p^{-1}\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*\top}) \le C$ and $\max_{1\le j\le p} \|\boldsymbol{\gamma}_j^*\|_2 \le C$.

Like all nonlinear (nonconvex) analyses, the rows of $\boldsymbol{B}^*$ and $\boldsymbol{\Gamma}^*$ are assumed to be in a bounded set, as in Assumptions 1 and 3. The boundedness of the natural parameter $\boldsymbol{\Theta}^*$ is required to control the tail probability of the response y conditional on observed covariates and latent factors. In Assumption 2, the sub-Gaussian assumptions admit the particular case when $x_{i1} = 1$ for all $1 \le i \le n$ so that the intercept can be incorporated into our model. In Assumption 3, the zero-mean condition on $\boldsymbol{W}^*$ is to simplify the theoretical analysis, which can be guaranteed if we include the intercept and project $\boldsymbol{Z}^*$ onto the linear span of the columns of $[\boldsymbol{1}_n, \boldsymbol{X}]$. Finally, the sparsity and boundedness assumptions of $\boldsymbol{B}^*$ in Assumption 1, and the bounded spectrum assumptions of $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Gamma}^*$ in Assumption 3 imply the conditions of Proposition 1 with $\tau_p = p$ therein. These assumptions are relatively lenient on the projection coefficient $\boldsymbol{D}^*$, provided they ensure that $\boldsymbol{\Theta}^*$ remains within a bounded set with high probability.

Remark 1 (The number of latent factors)**.** For our theoretical results, we assume the number of latent factors r is known in advance. Note that the joint-likelihood-based information criterion (JIC) proposed by Chen and Li [26] can be utilized to select the number of latent factors. The JIC value is the sum of deviance and a penalty on model complexity:

$$\mathrm{JIC}(\widehat{\boldsymbol{\Theta}}^{(r)}) = -2 \sum_{i\in[n], j\in[p]} \log\ p(y_{ij} \mid \widehat{\theta}_{ij}^{(r)}) + \nu(n, p, d + r), \tag{2.3.1}$$

where $\widehat{\boldsymbol{\Theta}}^{(r)}$ is the joint maximum likelihood estimator of the natural parameter matrix that minimizes (2.2.3) with r latent factors and d observed covariates, and $\nu(n, p, r) = c_{\mathrm{JIC}} \cdot r \log(n \wedge p)(n \wedge p)^{-1}$ is the complexity measure with penalty level $c_{\mathrm{JIC}} > 0$. As shown by Chen and Li [26], minimizing the empirical JIC yields a consistent estimate for the number of factors in generalized linear factor models with an intercept parameter. The utility of this metric in our problem setting is also empirically examined for both the simulation in Section 2.5 and the real data analysis in Section 2.6.

As motivated in Section 2.2.1, we consider the following optimization problem:

$$\begin{aligned} \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{\Gamma}} = & \operatorname*{argmin}_{\boldsymbol{B}\in\mathbb{R}^{p\times d}, \boldsymbol{Z}\in\mathbb{R}^{n\times r}, \boldsymbol{\Gamma}\in\mathbb{R}^{p\times r}} \mathcal{L}(\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{\Gamma}) + \lambda\|\boldsymbol{B}\|_{1,1} \\ & \text{s.t.} \qquad \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top \in \mathcal{R}_C^{n\times p}, \qquad \mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B} = \boldsymbol{0}. \end{aligned} \tag{2.3.2}$$

where the unregularized loss function $\mathcal{L}(\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{\Gamma})$ is defined in (2.2.3) and $\|\cdot\|_{1,1}$ denotes the element-wise ℓ_1-norm. It is worth noting that for any feasible $\widehat{\boldsymbol{\Gamma}}$ fixed, (2.3.2) reduces to a convex optimization problem in variables $\boldsymbol{B}$ and $\boldsymbol{Z}$:

$$\begin{aligned} \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}} = & \operatorname*{argmin}_{\boldsymbol{B}\in\mathbb{R}^{p\times d}, \boldsymbol{Z}\in\mathbb{R}^{n\times r}} \mathcal{L}(\boldsymbol{B}, \boldsymbol{Z}, \widehat{\boldsymbol{\Gamma}}) + \lambda\|\boldsymbol{B}\|_{1,1} \\ & \text{s.t.} \qquad \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}^\top \in \mathcal{R}_C^{n\times p}, \qquad \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{B} = \boldsymbol{0}. \end{aligned} \tag{2.3.3}$$

Algorithm 1 GCATE (generalized confounder adjustment for testing and estimation)

Input: A data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, a design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, a natural number $r \geq 1$ (the number of latent factors)

1: **Estimation of uncorrelated latent components $\boldsymbol{W}\boldsymbol{\Gamma}^\top$:** Solve optimization problem (2.3.4) to obtain $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top$ and the initial estimate $\widehat{\boldsymbol{\Theta}}_0 = \boldsymbol{X}\widehat{\boldsymbol{F}}^\top + \widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top$ by alternative maximization (Algorithm A.6.6) with initialization given in Appendix A.6.2:

$$\begin{aligned} \widehat{\boldsymbol{F}}, \widehat{\boldsymbol{W}}_0, \widehat{\boldsymbol{\Gamma}}_0 \in & \underset{\boldsymbol{F}\in\mathbb{R}^{p\times d}, \boldsymbol{W}\in\mathbb{R}^{n\times r}, \boldsymbol{\Gamma}\in\mathbb{R}^{p\times r}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}\boldsymbol{F}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top) \\ & \text{subject to} \quad \boldsymbol{X}\boldsymbol{F}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top \in \mathcal{R}_C^{n\times p}, \qquad \mathcal{P}_{\boldsymbol{X}}\boldsymbol{W} = \boldsymbol{0}. \end{aligned} \tag{2.3.4}$$

2: **Estimation of latent coefficients $\boldsymbol{\Gamma}$:** Set $\widehat{\boldsymbol{W}} := \sqrt{n}\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\widehat{\boldsymbol{\Gamma}} := \sqrt{p}\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$, where $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top = \sqrt{np}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the condensed SVD with $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, $\boldsymbol{V} \in \mathbb{R}^{p\times r}$.

3: **Estimation of direct effects $\boldsymbol{B}$ and latent factors $\boldsymbol{Z}$:** Solve optimization problem (2.3.5) to obtain $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}})$ by Algorithm A.6.6 with initialization given in Appendix A.6.2:

$$\begin{aligned} \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}} = & \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}, \boldsymbol{Z}\in\mathbb{R}^{p\times r}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}^\top) + \lambda\|\boldsymbol{B}\|_{1,1} \\ & \text{subject to} \quad \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}^\top \in \mathcal{R}_C^{n\times p}, \qquad \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{B} = \boldsymbol{0}. \end{aligned} \tag{2.3.5}$$

4: **Debiasing:** Construct the debiased estimate (2.4.1) and its estimated variance (2.4.6), based on $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{\Gamma}})$. Compute p-values according to the asymptotic distribution (2.4.5).

Output: Return the p-values.

This motivates us to solve optimization problem (2.3.2) in two steps: (1) firstly obtaining a good estimate of $\widehat{\boldsymbol{\Gamma}}$ and (2) then based on $\widehat{\boldsymbol{\Gamma}}$, obtaining good estimates for $\boldsymbol{B}^*$ and $\boldsymbol{Z}^*$. In Algorithm 1, we incorporate the two-step procedure by solving two sub-problems (2.3.4) and (2.3.5) consecutively. We next analyze the statistical properties of estimators in each step of Algorithm 1.

2.3.1 Estimation of uncorrelated latent components

To estimate the marginal effects $\boldsymbol{F}^* = \boldsymbol{B}^* + \boldsymbol{\Gamma}^*\boldsymbol{D}^*$ and the uncorrelated latent components $\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}$, we first solve optimization problem (2.3.4). This is also known as the joint maximum likelihood estimation [27, 28], which is statistically optimal in the minimax sense when both the sample size n and the response dimension p grow to infinity. From optimization problem (2.3.4), we obtain the initial estimates of the natural parameter matrix $\widehat{\boldsymbol{\Theta}}_0 = \boldsymbol{X}\widehat{\boldsymbol{F}}^\top + \widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top$. The following theorem characterizes the estimation error of the initial maximum likelihood estimate $\widehat{\boldsymbol{\Theta}}_0$.

Theorem 2 (Estimation error of $\widehat{\boldsymbol{\Theta}}_0$)**.** Under Assumptions 1–3, let $\widehat{\boldsymbol{\Theta}}_0$ be any estimator such that $\mathcal{L}(\widehat{\boldsymbol{\Theta}}_0) \leq \mathcal{L}(\boldsymbol{\Theta}^*)$. For any constant $\delta > 1$, when $np \geq 3$, it holds with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \iota_n$ that

$$\|\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|_{\mathrm{F}} = \mathcal{O}\left(\sqrt{(d+r)((n\vee p)\vee\delta^3)}\right), \ \max_{1\leq j\leq p}\|(\widehat{\boldsymbol{\Theta}}_0)_j - \boldsymbol{\Theta}_j^*\|_2 = \mathcal{O}\left(\sqrt{(d+r)(n\vee\delta^3)}\right).$$

In our specific setting, where the dimensions represented by d and r are orders of magnitude

smaller compared to n and p, the estimation error is primarily dominated by the scale of the larger dimensions n and p. Because the dimensions of natural parameters expand with both n and p, the bound implies that the error associated with each entry is approximately on the order of $\sqrt{(n \vee p)/np}$. As demonstrated in the upcoming subsection, these results empower us to attain robust estimates of the confounding effects.

2.3.2 Estimation of latent coefficients

From optimization problem (2.3.4), we obtain the initial estimates $\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{W}}_0 \widehat{\boldsymbol{\Gamma}}_0^\top$ of the latent components that are uncorrelated to the observed covariates. Though $\boldsymbol{W}^*$ and $\boldsymbol{\Gamma}^*$ are only identified up to rotations, we can use the condensed singular value decomposition of the normalized latent components $\widehat{\boldsymbol{E}}/\sqrt{np} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ to obtain the final estimates through the following optimization problem:

$$\begin{aligned} \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Gamma}} \in & \underset{\boldsymbol{W} \in \mathbb{R}^{n\times r}, \boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}}{\operatorname{argmin}} \frac{1}{np} \|\widehat{\boldsymbol{E}} - \boldsymbol{W}\boldsymbol{\Gamma}^\top\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \frac{1}{n}\boldsymbol{W}^\top \boldsymbol{W} = \frac{1}{p}\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma} \text{ is diagonal.} \end{aligned} \tag{2.3.6}$$

A simple derivation yields the solution $\widehat{\boldsymbol{W}} = \sqrt{n}\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\widehat{\boldsymbol{\Gamma}} = \sqrt{p}\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$ to the above problem. The above procedure is also commonly used in the factor analysis literature for estimating factor loadings from regression residuals [7, 20]. The following theorem guarantees that the above estimate of the latent coefficients is provably accurate in recovering the column space of the true coefficients.

Theorem 3 (Estimation error of $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}$)**.** Under Assumptions 1–3, as $n, p \to \infty$, it holds that

$$\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\boldsymbol{\Gamma}^*}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{1}{n \wedge p}}\right), \qquad \max_{1\le i,j\le p} |(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\boldsymbol{\Gamma}^*})_{ij}| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{1}{p^2(n \wedge p)}}\right).$$

Theorem 3 implies that the image of $\boldsymbol{\Gamma}^*$ can be estimated well by $\widehat{\boldsymbol{\Gamma}}$. Furthermore, the column-wise error decays at a fast rate. The precise error control of each column individually enables us to disentangle the intricate relationships within the confounder-adorned high-dimensional dataset. However, these do not directly extend to error control for the latent factors $\boldsymbol{Z}^*$ or the latent coefficients $\boldsymbol{\Gamma}^*$ themselves.

Under multivariate linear models, Bing et al. [20, Theorem 4] show the concentration of the latent coefficients $\max_{1\le j\le p} \|\boldsymbol{\gamma}_j^* - \widehat{\boldsymbol{\gamma}}_j\|_2$. Their results rely on the special structure of the regression problem $\boldsymbol{Y}_j = \boldsymbol{X}(\boldsymbol{b}_j^* + \boldsymbol{D}^{*\top}\boldsymbol{\gamma}_j^*) + \boldsymbol{\epsilon}_j$, where the regression coefficient can be decomposed into a sparse component $\boldsymbol{b}_j^*$ and a dense component $\boldsymbol{D}^{*\top}\boldsymbol{\gamma}_j^*$. By using the lava estimator [31], they can derive the estimation error of the residual $\boldsymbol{\epsilon}_j$, from whose covariance structure, $\|\boldsymbol{\gamma}_j^* - \widehat{\boldsymbol{\gamma}}_j\|_2$ can be further bounded. In the generalized linear model setting, we don't have the flexibility to utilize the additive noises' covariance structure to directly estimate $\boldsymbol{\gamma}_j$ well. Instead, we need to rely on joint maximum likelihood estimation to estimate them, as we illustrate below.

2.3.3 Estimation of latent factors and direct effects

Once the column space of confounding effects becomes distinguishable, the subsequent phase entails retrieving direct effects by mitigating the influence of confounding variables and solving

optimization problem (2.3.3). Through this, we can simultaneously obtain the estimates of latent factors $\boldsymbol{Z}^*$ and direct effects $\boldsymbol{B}^*$. In the high-dimensional scenarios when p can be larger than n, one natural approach to estimate the sparse coefficient $\boldsymbol{B}^*$ is via ℓ_1-regularization, as employed in the optimization problem (2.3.5), which aims to obtain a sparse and consistent estimator $\widehat{\boldsymbol{B}}$ by ℓ_1-regularization while simultaneously removing the unmeasured effects. Next, we analyze the properties of the two estimators $\widehat{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{B}}$ in turn.

Latent factors. As previously alluded to, estimating latent factor $\boldsymbol{Z}^*$ demands special consideration. To bypass the technical difficulty, we will use the estimation errors of $\widehat{\boldsymbol{\Theta}}_0$ and $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}$ provided by Theorem 2 and Theorem 3, respectively, coupled with one extra identifiability condition for the latent factors $\boldsymbol{Z}^*$ and their coefficients $\boldsymbol{\Gamma}^*$. From Lin et al. [104, Proposition 5.1], there exists an invertible matrix $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ with bounded operator norm, such that $\mathbb{V}(\boldsymbol{R}\boldsymbol{z}_1^*)$ and $\boldsymbol{R}^{-\top}\boldsymbol{\gamma}_1^*\boldsymbol{\gamma}_1^{*\top}\boldsymbol{R}^{-1}$ are the same diagonal matrix. The following assumption from Lin et al. [104] restricts the spacing of $\mathbb{V}(\boldsymbol{R}\boldsymbol{z}_1^*)$'s eigenvalues.

Assumption 4 (Identifiability of latent factors)**.** Assume there exists positive numbers $c_1 \le c_2$ and $1 < k_1 \le k_2$ such that for all $\ell \in \{1,\dots,r\}$, the eigenvalues of $\mathbb{V}(\boldsymbol{R}\boldsymbol{z}_1^*)$ satisfy $c_1\ell^{-k_1} \le \lambda_\ell \le c_2\ell^{-k_1}$, and $\lambda_\ell - \lambda_{\ell+1} \ge c_1\ell^{-k_2}$, with the convention that $\lambda_{r+1} = 0$.

Intuitively, Assumption 4 guarantees that $\boldsymbol{Z}^*\boldsymbol{R}^\top$ can be recovered up to sign from the matrix product $\boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}$. This implies that, if one can consistently estimate $\boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}$ with $\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Gamma}}^\top$, then $\boldsymbol{Z}^*$ and $\boldsymbol{\Gamma}^*$ can also be consistently estimated by appropriate transformations of $\widehat{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{\Gamma}}$, respectively. A simple consequence from Lin et al. [104, Proposition 5.2], coupled with Theorems 2 and 3, is the following error bound on the columns of latent components.

Corollary 4 (Estimation of latent components)**.** Let $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{Z}}$ be solutions to optimization problems (2.3.6) and (2.3.5), respectively. Under Assumptions 1–4, suppose $\min_{\{\boldsymbol{B}\in\mathbb{R}^{p\times d}|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{B}=\boldsymbol{0}\}} \mathcal{L}(\boldsymbol{X}\boldsymbol{B}^\top + \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Gamma}}^\top) \le \mathcal{L}(\boldsymbol{\Theta}^*)$ with probability tending to one. Then, as $n, p \to \infty$, it holds that,

$$\max_{1\le j\le p} \frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{Z}^*\boldsymbol{\gamma}_j^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}} \vee \sqrt{\frac{r^{4k_2-k_1+4}\log n}{n\wedge p}}\right).$$

It's important to note that, unlike the analysis for linear models in prior work [19, 20] that projects the responses onto the orthogonal column space of $\widehat{\boldsymbol{\Gamma}}$ and removes the effects of latent factors $\boldsymbol{Z}$, estimating $\boldsymbol{Z}$ is unavoidable under generalized linear models. In Corollary 4, we require the joint maximum likelihood based on the estimated latent components to be higher than the likelihood evaluated at the truth. This requires the estimated latent components derived from (2.3.5) to exhibit stability and ensures that the maximum likelihood with the estimated latent factors remains close to the joint maximum likelihood from (2.3.4). In the presence of nuisance parameters $\boldsymbol{Z}^*$ and $\boldsymbol{\Gamma}^*$, the sharp control on estimation error of the column $\boldsymbol{Z}^*\boldsymbol{\gamma}_j^*$ provided by Corollary 4 helps control the estimation error of $\widehat{\boldsymbol{B}}$.

Direct effects. In high-dimensional scenarios, controlling the estimation error of $\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*$ requires the projection $\mathcal{P}_{\boldsymbol{\Gamma}^*}$ does not excessively densify the primary effects $\boldsymbol{B}^*$. To this end, we require the ratio $\|\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*\|_{1,1}/\|\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*\|_{\mathrm{F}}$ to be of smaller order than $\sqrt{p}$. Coupled with the previous assumptions and results, the estimation error of $\widehat{\boldsymbol{B}}$ returned by problem (2.3.5) can be controlled.

Theorem 5 (Estimation error of $\widehat{\boldsymbol{B}}$)**.** Suppose the assumptions in Corollary 4 hold and $\|\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*\|_{1,1} = \mathcal{O}(p^{k/2}\|\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*\|_{\mathrm{F}})$ for some constant $k \in [0,1)$. Then, as $n, p \to \infty$ such that $\sqrt{n}/\log(nd) = o(p)$ and $\log(p) = o(n)$, the estimate $\widehat{\boldsymbol{B}}$ of optimization problem (2.3.5) with $\lambda \asymp 8\nu^2\log(nd)n^{-1/2}$

satisfies that

$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{\mathrm{F}} = \mathcal{O}_{\mathbb{P}}(r_{n,p}), \qquad \|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{1,1} = \mathcal{O}_{\mathbb{P}}(r'_{n,p}),$$

where the sequences $r_{n,p}$ and $r'_{n,p}$ are defined as:

$$r_{n,p} := \sqrt{\frac{(sd\log^2(nd)) \vee \log(np)}{n}} + \sqrt{\frac{n^{1/2}}{(n \wedge p)^{3/2}\log(nd)}} + \sqrt{\frac{sd}{n \wedge p^{1-k}}},$$
$$r'_{n,p} := \sqrt{sd}\, r_{n,p} + \frac{\sqrt{n}}{(n \wedge p)\log(nd)}.$$

In Theorem 5, the parameter k captures the deviation of the projected ℓ_1-norm $\|\mathcal{P}^{\perp}_{\boldsymbol{\Gamma}^*}\boldsymbol{B}^*\|_{1,1}$ from $\|\boldsymbol{B}^*\|_{1,1}$. The smaller k, the more information of $\boldsymbol{B}^*$ is retained after projection, and the signal-noise-ratio is larger. In the high-dimensional scenarios when $n < p$, the ℓ_1-norm and ℓ_2-norm of the estimation error scale in $\mathcal{O}_{\mathbb{P}}(n^{-1/2} \wedge p^{(k-1)/2})$ when ignoring lower order factors. The appearance of the response dimension p in the denominator reflects the blessing of dimensionality; namely, having more responses than the sample size is not detrimental to consistency, provided that p does not grow exponentially larger than n, as we numerically demonstrate in Figure A.73.

To prove Theorem 5, we establish the (approximate) optimal condition for $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{\Gamma}})$ from the two-step procedure to the joint optimization problem (2.3.2), as shown in Lemma A.4.1. This relies on the optimality condition of optimization problem (2.3.5) and the convergence rate of $\mathcal{P}^{\perp}_{\widehat{\boldsymbol{\Gamma}}}$ provided by Theorem 3. It then allows us to establish the cone condition, obtain the upper and lower bounds of the first-order approximation error of the loss function, and derive the error rate in Appendix A.4. Compared to double machine learning in the presence of high-dimensional nuisance parameters [32, 33], our estimation procedure does not require sample splitting. To establish consistency, we only need the convergence rate of $\max_{1\le j\le p}\|\boldsymbol{Z}^*\boldsymbol{\gamma}_j^* - \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}_j\|_2/\sqrt{n}$ to be the parametric rate $(n\wedge p)^{-1/2}$, as shown in the proof of Theorem 5; see also Remark 12 for discussion on the connection to Neyman orthogonality. However, to derive the asymptotic distribution for inference, one may need more stringent conditions or sample splitting, as illustrated next.

2.4 Inference

2.4.1 Projected and weighted bias correction

When evaluating uncertainty in high-dimensional inference, confidence intervals and statistical hypothesis tests are required. After obtaining the initial estimate $\widehat{\boldsymbol{B}}$, we need to remove the bias caused by ℓ_1-regularization to have valid inferences on the estimated coefficients. Without loss of generality, we focus on testing the coefficients of the first covariate b_{j1} for $j = 1, \ldots, p$. We consider the following debiased estimator for each of them:

$$\widehat{b}^{\mathrm{de}}_{j1} = \widehat{b}_{j1} + \boldsymbol{u}^{\top}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i))^{\top}\boldsymbol{v}_i, \tag{2.4.1}$$

where $\widehat{\boldsymbol{\Theta}} := \boldsymbol{X}\widehat{\boldsymbol{B}}^{\top} + \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Gamma}}^{\top}$ is the estimated natural parameter matrix, and $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v}_i \in \mathbb{R}^p$ are projection vectors to be specified later, such that the correction term $n^{-1}\boldsymbol{u}^{\top}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i))^{\top}\boldsymbol{v}_i$ is a reasonable estimate of the bias $b^*_{j1} - \widehat{b}_{j1}$.

By Taylor expansion of $A'(\theta_{ij}^*)$ at $\widehat{\theta}_{ij} := \boldsymbol{x}_i^\top \widehat{\boldsymbol{b}}_j + \widehat{\boldsymbol{z}}_i^\top \widehat{\boldsymbol{\gamma}}_j$, we have

$$A'(\theta_{ij}^*) = A'(\widehat{\theta}_{ij}) + A''(\widehat{\theta}_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij}) + \frac{1}{2}A'''(\psi_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij})^2,$$

for some ψ_{ij} between $\widehat{\theta}_{ij}$ and θ_{ij}^*. Then, the residual of the ith sample can be decomposed into three sources of errors:

$$\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i) = \underbrace{\boldsymbol{\epsilon}_i}_{\text{stochastic error}} + \underbrace{\boldsymbol{p}_i}_{\text{remaining bias}} + \underbrace{\boldsymbol{q}_i}_{\text{approximation error}} \tag{2.4.2}$$

where the three terms of errors are given by

$$\begin{aligned}
\boldsymbol{\epsilon}_i &= \boldsymbol{y}_i - A'(\boldsymbol{\theta}_i^*) \\
\boldsymbol{p}_i &= A''(\widehat{\boldsymbol{\theta}}_i) \odot (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*) \\
\boldsymbol{q}_i &= -\frac{1}{2}[A'''(\psi_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij})^2]_{1\le j\le p}.
\end{aligned}$$

If we let $\boldsymbol{v}_i = \omega_i \mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j$ where ω_i is the sample-specific weight and $\boldsymbol{e}_j \in \mathbb{R}^p$ is the unit vector with jth entry being one, then substituting (2.4.2) into (2.4.1) yields that

$$\begin{aligned}
\widehat{b}_{j1}^{\mathrm{de}} - b_{j1}^* &= (\widehat{b}_{j1} - b_{j1}^*) + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top \boldsymbol{v}_i + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{p}_i^\top \boldsymbol{v}_i + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{q}_i^\top \boldsymbol{v}_i \\
&= \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top \boldsymbol{v}_i + \left(\boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{e}_1^\top\right)(\boldsymbol{b}_j^* - \widehat{\boldsymbol{b}}_j) + \mathrm{Rem}.
\end{aligned} \tag{2.4.3}$$

The estimation error rates provided by Theorems 3 and 5 guarantee that $\|\boldsymbol{b}_j^* - \widehat{\boldsymbol{b}}_j\|_1 = \mathcal{O}_{\mathbb{P}}(r'_{n,p})$ and the remaining term is $\mathrm{Rem} = \mathcal{O}_{\mathbb{P}}(\max_{1\le i\le n}|\boldsymbol{u}^\top\boldsymbol{x}_i|^3 r_{n,p}^2)$ for $r_{n,p}$ and $r'_{n,p}$ defined in Theorem 5. Based on (2.4.3), the idea of debiasing is to choose $\boldsymbol{u}$ and ω_i's such that the second term and the remaining term is of order $o_{\mathbb{P}}(n^{-1/2})$, while enabling the convergence of the average of primary stochastic errors to a normal distribution by central limit theorem.

To facilitate our theoretical analysis, suppose we split the dataset into two parts $\mathcal{D}_1 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i), 1 \le i \le n\}$ and $\mathcal{D}_2 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i), n+1 \le i \le 2n\}$, where $\mathcal{D}_2$ is used to obtain the estimates $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Gamma}}$, and $\mathcal{D}_1$ is used to remove the bias for $\widehat{\boldsymbol{B}}$ induced by ℓ_1-regularization. There are also latent factors $\{\boldsymbol{z}_i^*\}_{i=1}^n$ and $\{\boldsymbol{z}_i^*\}_{i=n+1}^{2n}$ associated with $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Further, if $\boldsymbol{\epsilon}_i$'s are independent of the projection vectors $\boldsymbol{u}$ and $\boldsymbol{v}_i$ (or equivalently ω_i) conditional on $(\boldsymbol{x}_i, \boldsymbol{z}_i^*)$'s and $\mathcal{D}_2$, then we can approximate the scaled conditional variance of the stochastic errors as:

$$\begin{aligned}
\sigma_j^2 &= \mathbb{V}\left(\boldsymbol{u}^\top \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top \boldsymbol{v}_i \,\middle|\, \{(\boldsymbol{x}_i, \boldsymbol{z}_i^*)\}_{i=1}^n, \mathcal{D}_2\right) \\
&= \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i^2 \boldsymbol{x}_i\boldsymbol{e}_j^\top \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\,\mathrm{Cov}(\boldsymbol{\epsilon}_i \mid \boldsymbol{\theta}_i^*)\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j\boldsymbol{x}_i\boldsymbol{u} \\
&= \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i^2 (\boldsymbol{e}_j^\top \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{u} \\
&\approx \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{u} =: \widehat{\sigma}_j^2,
\end{aligned}$$

by using a proper data-dependent weight $\omega_i = \widehat{\omega}_i$. Then, the projection vector $\widehat{\boldsymbol{u}}$ is constructed by minimizing the variance proxy while controlling the bias and remaining terms in (2.4.3):

$$\begin{aligned} \widehat{\boldsymbol{u}} \quad \in \quad & \underset{\boldsymbol{u}\in\mathbb{R}^d}{\operatorname{argmin}}\, \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \widehat{\omega}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{u} \\ \text{s.t.} \quad & \left\| \frac{1}{n}\sum_{i=1}^n \widehat{\omega}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{u} - \boldsymbol{e}_1 \right\|_\infty \le \lambda_n \\ & \max_{1\le i\le n} |\boldsymbol{x}_i^\top \boldsymbol{u}| \le \tau_n, \end{aligned} \tag{2.4.4}$$

where $\lambda_n \asymp \sqrt{\log(nd)/n}$ and $\tau_n \asymp \sqrt{\log n}$. Based on $\widehat{\omega}_i$ and $\widehat{\boldsymbol{u}}$, the resulting bias-corrected estimator (2.4.1) is similar to those used by Cai et al. [21], Javanmard and Montanari [74], van de Geer et al. [168]; however, we need to incorporate information from multiple responses with projection operator $\mathcal{P}^{\perp}_{\widehat{\boldsymbol{\Gamma}}}$ to de-confound, in the spirit of proximal gradient descent. Under mild regularity conditions, the following theorem shows that the debiased estimator $\widehat{b}^{\mathrm{de}}_{j1}$ is asymptotically normal.

Theorem 6 (Asymptotical normality of $\widehat{\boldsymbol{B}}^{\mathrm{de}}$)**.** Under the same conditions in Theorem 5, for $j = 1, \ldots, p$, additionally assume the following conditions hold: (i) $n/\log(nd) = o(p^{3/2})$ and $n = o(p^{2(1-k)})$; and (ii) $\widehat{\omega}_i = \omega(\boldsymbol{x}_i, \boldsymbol{z}_i^*, \mathcal{D}_2)$ for some real-valued function ω that is uniformly bounded away from 0 and ∞. Then as $n, p \to \infty$, it holds that

$$\sqrt{n}\frac{\widehat{b}^{\mathrm{de}}_{j1} - b^*_{j1}}{\sigma_j} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{2.4.5}$$

With fewer assumptions on the correlation between the covariate $\boldsymbol{X}$ and the confounder $\boldsymbol{Z}^*$, removing unmeasured confounders is only possible by utilizing multiple outcomes to disentangle the primary effect $\boldsymbol{B}^*$ and the latent coefficient $\boldsymbol{\Gamma}^*$. In particular, because the estimation error rates of $\boldsymbol{B}^*$ and $\boldsymbol{\Gamma}^*$ are related to $(n \wedge p)^{-1}$, the number of outcomes p is expected to be larger than n, so that these errors are primarily affected by the sample size n.

In Theorem 6, condition (i) requires that the response dimension p grows faster than $n^{2/3} \vee n^{1/(2(1-k))}$, which ensures the remainder term Rem in (2.4.3) vanishes in the limit. Specifically, Rem has a magnitude associated with the convergence rate of $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{\mathrm{F}}^2$, as provided by Theorem 5. To derive the asymptotic normality, Rem $= o_{\mathbb{P}}(n^{-1/2})$ is required; however, if n is too large compared to p, the convergence rate of $\widehat{\boldsymbol{B}}$ from the first two steps of the proposed procedure is insufficient to establish the desired asymptotic normality. In this case, having a much larger sample size does not help. When $k \le 1/2$, condition (i) is satisfied with $n = o(p)$, which is reasonable in most scientific scenarios of cohort-level differential expression analysis, as we shall see later from the real data example in Section 2.6.

In terms of condition (ii), a proper sample-specific and link-specific weight function is required. One can construct such weights $\widehat{\omega}_i$ by sample splitting to fulfill this condition. For instance, using sample splitting procedure in Algorithm A.5.5, one valid choice is $\widehat{\omega}_i = A''(\widehat{\theta}_{ij})$. In Lemma A.5.3, we show that such a choice of $\widehat{\omega}_i$ satisfies the condition (ii) in Theorem 6 with probability tending to one and the resulting variance estimator

$$\widehat{\sigma}_j^2 = \widehat{\boldsymbol{u}}^\top \frac{1}{n}\sum_{i=1}^n \widehat{\omega}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{u}}, \tag{2.4.6}$$

is also consistent with σ_j^2. Hence, Theorem 6 implies that $t_j = \sqrt{n}(\widehat{b}_{j1}^{\mathrm{de}} - b_{j1}^*)/\widehat{\sigma}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. We reject the null hypothesis $H_{0j} : b_{j1}^* = 0$ at level-α if $|t_j| > z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$, where Φ is the cumulative distribution function of standard normal. Numerically, we show that the efficiency loss of sample splitting is negligible in Appendix A.7.1; and the proposed method performs well without sample splitting and is statistically more efficient than the alternative methods in Section 2.5.

Remark 2 (Inference without unmeasured confounders)**.** In the special case when there are no unmeasured confounders, the matrix $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}$ reduces to the identity matrix. Also, the projection vector $\widehat{\boldsymbol{u}}$ is the jth column of $(\boldsymbol{X}^\top \mathrm{diag}(A''(\boldsymbol{X}\widehat{\boldsymbol{B}}_1))\boldsymbol{X})^{-1}$, and $\widehat{\sigma}_j$ is the asymptotic variance of $\widehat{b}_{j1}$ under well-specified generalized linear models. In this case, (2.4.1) is simply a one-step adjustment based on the score function. For Bernoulli distributed binary outcomes without unmeasured confounders, the above choice of the weight function $\omega(\theta) = A''(\theta)$ for optimization problem (2.4.4) coincide with $f'(\theta)^2/(f(\theta)(1-f(\theta)))$, the one used in Cai et al. [21] with $f = A'$ being the link function. For Gaussian outcomes when A' is the identity link with a choice of weight $\widehat{\omega}_i \equiv 1$, the procedure above reduces to the debias method by Javanmard and Montanari [74].

When there are unmeasured confounders, the main difficulty lies in taking account of the rates of convergence of $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_1$, $\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\boldsymbol{e}_j\|_2$, and $\|\boldsymbol{B}^{*\top}\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{e}_j\|_2$ such that the remainder term in (2.4.3) is $o_{\mathbb{P}}(n^{-1/2})$, which is essentially the idea of the proof for Theorem 6, as we have alluded to after (2.4.3).

Remark 3 (Incorporate information from latent factors)**.** In (2.4.1) and (2.4.4), we only use the covariate $\boldsymbol{X}$ to adjust for the estimation bias. However, including the estimated latent factors $\boldsymbol{Z}$ to construct a projection vector $\boldsymbol{u}$ of dimension $d + r$ is also feasible. The validity of this extension is also guaranteed by the sample splitting procedure in Algorithm A.5.5.

Remark 4 (Estimation and inference with non-canonical links)**.** Through Sections 2.3 and 2.4, we discuss the methodology to conduct inference on confounded generalized linear models (GLM) with canonical link functions, as outlined in Table A.61. However, in practical scenarios, non-canonical link functions may also be employed. For instance, the log link function is commonly used with Negative Binomial GLMs. Fortunately, our method extends its applicability to GLMs with non-canonical link functions, as exemplified in the case of the Negative Binomial GLMs in Appendix A.6.4. Establishing theoretical guarantees for these scenarios may follow a similar framework with suitable assumptions to address the non-convexity of the objective functions, as elaborated in Appendix A.6.4.

2.4.2 Simultaneous inference

For $j = 1, \ldots, p$, the asymptotic normality provided in Theorem 6 provides Type-I error controls for individual hypothesis tests $H_{0j} : b_{j1}^* = 0$. The following proposition shows that we can also control the overall Type-I error and family-wise error rate (FWER) using the statistic $t_j = \sqrt{n}(\widehat{b}_{j1}^{\mathrm{de}} - b_{j1}^*)/\widehat{\sigma}_j$.

Proposition 7 (Simultaneous inference)**.** Let $\mathcal{N}_p = \{j \mid b_{j1}^* = 0, j = 1, \ldots, p\}$ be the true null hypotheses. Under the assumptions of Theorem 6, as $n, p, |\mathcal{N}_p| \to \infty$, it holds that

$$\frac{1}{|\mathcal{N}_p|}\sum_{j\in\mathcal{N}_p} \mathbb{1}\{|t_j| > z_{\frac{\alpha}{2}}\} \xrightarrow{\mathrm{p}} \alpha, \qquad \text{and} \qquad \limsup\ \mathbb{P}\left(\sum_{j\in\mathcal{N}_p} \mathbb{1}\{|t_j| > z_{\frac{\alpha}{2p}}\} \geq 1\right) \leq \alpha.$$

When p is large, controlling for the false discovery rate (FDR) is more desirable when performing simultaneous testing. In that regard, Cai et al. [21, Section 2.3] provides insights on FDR controls using different techniques. From simulations in Section 2.5, we also show that FDR is usually well controlled by the Benjamini–Hochberg procedure empirically.

2.5 Numerical experiments

DE and related tests are frequently performed in two distinct settings in the genomic field. One relies on counts of gene expression to contrast the expression of each gene in case versus control observations. Typically, observations are either samples from RNA sequencing (RNA-seq) [109] or pseudo-bulk cells obtained from single-cell sequencing by aggregating the expressions of single cells in the same homogeneous groups [158]. Another setting is single-cell RNA-sequence (scRNA-seq) CRISPR screening [12, 37], where the fundamental task is to test for association between a designed genetic perturbation and gene expression [37]. In both settings, the measured gene expression is often assumed to approximately follow a Poisson or Negative Binomial (NB) distribution [150]. However, in the former, the mean expression per sample is much larger due to molecular design, and the distribution is often approximated by a normal distribution with an appropriate transformation. In the latter case, the observational unit is a single cell. Hence, the mean of the gene expression is near zero, and the data is not well approximated with a normal distribution.

Before we turn to the simulation details, we present a simulated bulk-cell dataset and a simulated single-cell dataset corresponding to the above two distinct scenarios, respectively (Figure 2.2). The Poisson distribution can often model the former scenario, while the NB distribution is a better option for the latter because the counts are sparser and typically exhibit strong overdispersion (Figure 2.2(a)-(b)). Furthermore, for single-cell data, the lower-expressed genes are typically more dispersed, and this feature is captured in our simulated data set (Figure 2.2(c)). In practice, both Poisson and NB models are available for analysis of either type of experiment; however, to simplify exposition, we use a Poisson distribution for bulk samples in Section 2.5.1 and a NB distribution for single-cell samples in Section 2.5.2. In the subsequent experiments, we adhere to the protocol described in Appendix A.6.3 for selecting both the hyperparameters and the number of factors pertinent to the proposed methods.

2.5.1 Well-specified simulated datasets

We simulate expression data $\boldsymbol{Y}$ that consists of $n \in \{100, 250\}$ cells and $p = 3,000$ genes based on the Poisson likelihood with natural parameter $\boldsymbol{\Theta}$. More specifically, we generate the covariate x_1 to be a centered binary variable, i.e., $(x_1+1)/2 \sim \text{Bernoulli}(0.5)$. We also include an intercept $x_2 = 1$, so that the covariate vector $\boldsymbol{x} = [x_1, x_2]^\top$ has dimension $d = 2$. To allow for the most general confounding scenarios without assuming causal relationships as in Figure 2.1, we directly generate the latent factor matrix using $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{D}^\top + \boldsymbol{W} \in \mathbb{R}^{n\times r}$ with the number of latent factors being $r \in \{2, 10\}$. Here, to generate $\boldsymbol{D}$ and $\boldsymbol{W}$, we first sample their entries independently from $\mathcal{N}(0, 1)$ and further modify the singular values to be $s_1, \ldots, s_r$ where $s_k = a \cdot (2-(k-1)/(r-1))$, with $a = n^{-3/2}$ for $\boldsymbol{D}$ and $a = (n/2)^{1/2}$ for $\boldsymbol{W}$. For the latent loading matrix $\boldsymbol{\Gamma}$, we follow Wang et al. [175] to take $\boldsymbol{\Gamma} = \widetilde{\boldsymbol{\Gamma}}\boldsymbol{\Lambda}$ where $\widetilde{\boldsymbol{\Gamma}}$ is a $p\times r$ orthogonal matrix sampled uniformly from the set of all $p\times r$ orthogonal matrix and $\boldsymbol{\Lambda} = (p/2)^{1/2}\text{diag}(\lambda_1, \ldots, \lambda_r)$ where $\lambda_k = 2-(k-1)/(r-1)$. The primary effect of x_1 on gene j is sampled from $(b_{j1} + 0.2)/0.4 \sim \text{Bernoulli}(0.5)$ with probability 0.05 and set to be zero with probability 0.95. The coefficient for the intercept is set to be $b_{j2} = 0.5$.

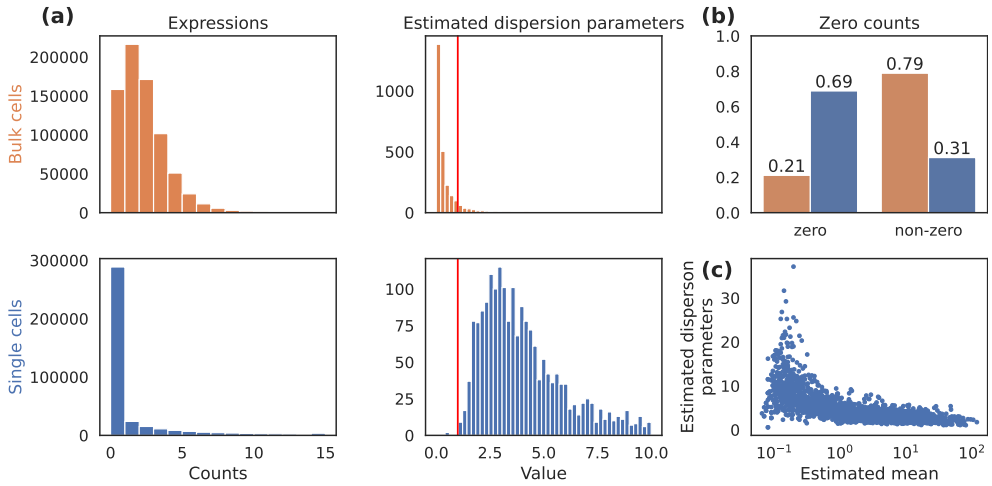

Figure 2.2: Overview of the simulated data. **(a)** The first and second rows show the summary of one simulated dataset for bulk cells (Poisson) in Section 2.5.1 and single cells (Negative Binomial) by Splatter in Section 2.5.2, respectively. The first column shows the overall distribution of the generated counts; the second column shows the estimated dispersion parameters by methods of moments using the mean estimates from GLM with Poisson likelihood. **(b)** The proportions of zero and non-zero counts in the two datasets, colored in orange and blue, respectively. **(c)** The estimated dispersion parameter versus the estimated mean for the simulated single-cell dataset.

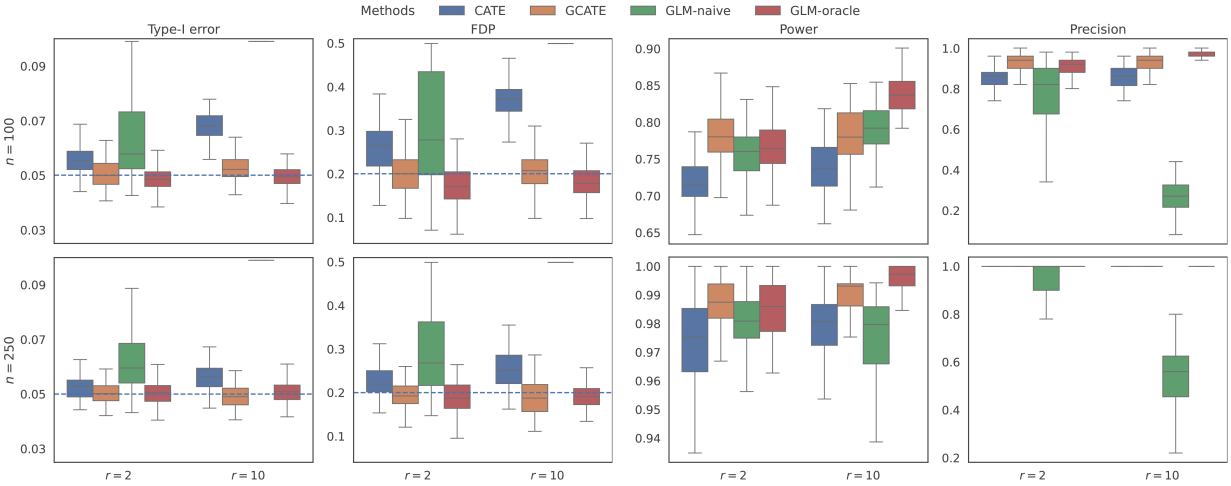

Figure 2.3: The Type-I errors, false discovery proportions (FDPs), powers, and precision of different methods on the simulated datasets over 100 runs, with varying numbers of samples $n \in \{100, 250\}$ and numbers of latent factors $r \in \{2, 10\}$. For GLM, the maximum values of Type-I errors and FDPs are clipped at 0.1 and 0.5, respectively. The blue dashed lines indicate the desired cutoffs.

Four methods are applied to the simulated datasets: (1) CATE (confounder adjustment for testing and estimation), which is a unified approach for surrogate variable analysis under linear models [175] and operates on the log-normalized data. (2) GLM-naive, which fits generalized linear models with the Poisson likelihood method but only uses the measured covariates $\boldsymbol{X}$ without adjusting for unmeasured confounders; (3) GLM-oracle, which fits generalized linear models with

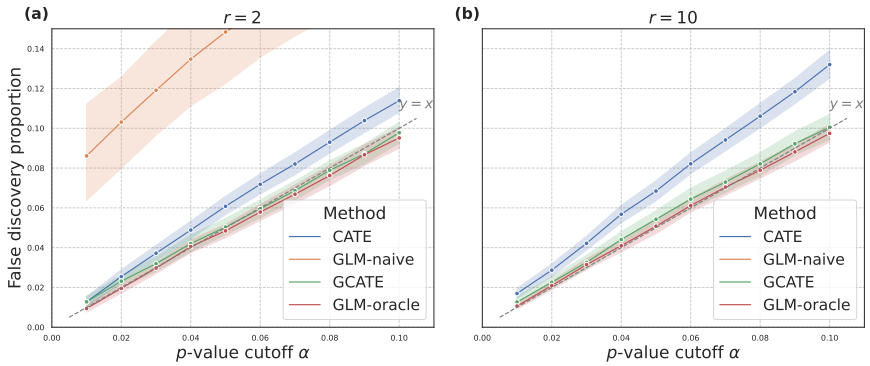

Figure 2.4: False discovery proportion at different α levels for p-values adjusted by the Benjamini-Hochberg procedure on 100 simulated datasets when $n = 250$. The left and right panels show the results for different numbers of latent factors, **(a)** $r = 2$ and **(b)** $r = 10$, respectively. When $r = 10$, the FDP of GLM-naive is above 0.15; hence it is not shown in the figure.

the Poisson likelihood method and uses both observed and unobserved covariates $(\boldsymbol{X}, \boldsymbol{Z})$ for estimation and testing; (4) GCATE, our proposed method with the Poisson likelihood. For CATE, we use bi-cross-validation (BCV) [132] to select the number of factors, as suggested in their original paper [175]. For GCATE, we use JIC described in Remark 1 to select the number of factors.

To evaluate different methods, we summarize the type-I error and FDP (false discovery proportion) after the Benjamini-Hochberg procedure in Figure 2.3, where the desirable thresholds for the two are set to be 5% and 20%, respectively. From Figure 2.3, we see that when the sample size n is small, or the latent factor dimension r is large, the performance of all methods gets slightly worse, especially those that are misspecified, which is expected. In all setups of (n, r), because the multivariate-Gaussian assumptions of CATE are violated, it does not provide proper Type-I error control and FDP control. This suggests that CATE may inflate test statistics and cause anti-conservative inference. Similarly, GLM-naive also fails to control the FDPs because it cannot account for dependencies induced by the latent factors. On the other hand, GCATE performs as well as GLM-oracle that has knowledge of the latent factors $\boldsymbol{Z}$. This indicates that our modeling helps to accurately remove unwarranted sources of confounding effects. Note that variations of CATE may yield improved performance using empirical nulls or negative controls, but GCATE requires no such tuning.

We further inspect the FDP control of different methods with varying thresholds. In the ideal scenarios, FDP aligns closely with the specified α cutoffs. From Figure 2.4, GLM-oracle has FDP aligning closely with the specified α cutoffs and consistently performs admirably across different levels of confounding effects. Conversely, the GLM-naive approach struggles to control the FDP effectively, and this discrepancy becomes increasingly pronounced as the number of latent factors grows. However, in a commendable contrast to CATE, our method GCATE consistently outperforms in terms of FDP control at various alpha cutoffs. This superiority can be attributed to our method's ability to model the data distribution accurately and eliminate unwarranted variations.

Lastly, we also evaluate the statistical power and precision of different methods. Here, the power is evaluated when the Type-I error threshold is 5%. We anticipate that both CATE and the GLM-naive approach would yield higher power compared to other methods because they tend to allow more discoveries without adequately controlling the Type-I errors (Figure 2.3). In Fig-

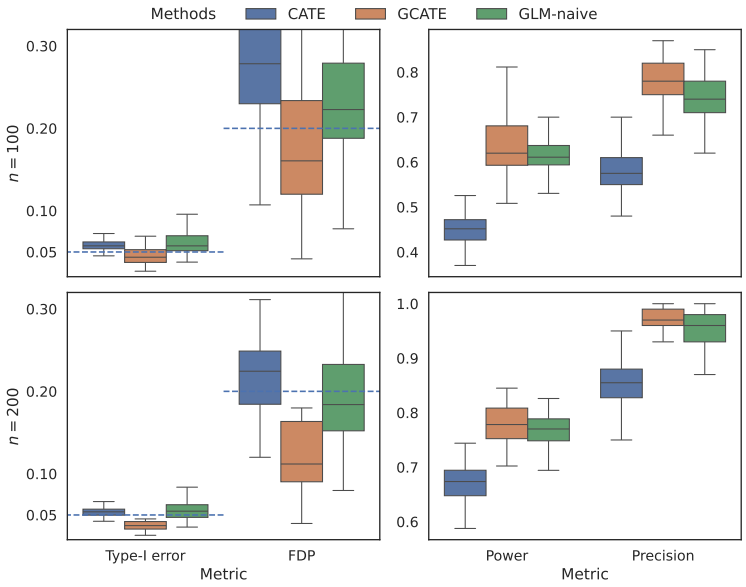

Figure 2.5: Simulation results on 100 simulated scRNA-seq datasets generated by Splatter with varying numbers of samples $n \in \{100, 200\}$. The four metrics are shown in four columns respectively. The blue dashed lines indicate the desired cutoffs for the statistical errors.

ure 2.3, we observe that CATE exhibits the lowest power among the considered methods. In contrast, the GLM-naive approach concurrently registers the most insufficient precision among all the methods. As anticipated, the GLM-oracle approach boasts the highest power and precision because it operates in an ideal scenario without confounding effects. In contrast, our proposed method, GCATE, demonstrates a balanced and robust performance concerning power and precision. It achieves a competitive power level while maintaining a significantly higher precision than the GLM-naive method. Moreover, GCATE outperforms CATE regarding both power and precision. This suggests that correct modeling of confounding effects boosts the statistical power and precision in high-dimensional datasets.

2.5.2 Misspecified simulated datasets using scRNA simulators

To better evaluate the performance of various methods, we use the single-cell RNA sequencing data simulator Splatter [182] to generate simulated count datasets. Splatter explicitly models the hierarchical Gamma-Poisson processes that give rise to data observed in scRNA-seq experiments and can model the multiple-faceted variability. Thus, the simulated datasets generated by Splatter are similar to real-world datasets and suitable for benchmarking differential expression testing methods.

Using Splatter, n cells are sampled from two groups with equal probability for $n \in \{100, 200\}$, containing $p = 10,000$ genes. Because of the sparse nature of the simulated single-cell datasets, about 80% of the genes are only expressed in 10 cells. Hence, we exclude these lowly-expressed genes and evaluate the methods for the remaining genes. We include $d = 3$ covariates for each cell: the intercept, the group indicator ($\{\pm 1\}$), and the logarithm of the library sizes, which is the sum of expression across all genes. When simulating the datasets, we use Splatter to generate four batches, introducing three major confounders. Because the data is not generated

from well-specified GLMs, the oracle model is unknown and hence not included. For GLM-naive and GCATE, we use the NB likelihood with log links to directly model the count data, where the gene-level dispersion parameters are estimated by the method of moments based on the estimated mean returned by using the Poisson likelihood; see Appendix A.6.2 for more details. For CATE, we normalize the counts in each cell by its library size, then multiply them by a scale factor of 10^4 and shift them by one, and finally, apply the logarithm transform, following the standard preprocessing approach of single-cell data.

Compared to the previous bulk-cell simulation in Section 2.5.1, the simulated data from Splatter is sparser and more noisy. From Figure 2.5, both CATE and GLM fail to control the Type-I error at level 5% and have lower power than GCATE in this more challenging setting. The primary reason lies in the assumption underlying CATE is significantly violated, while the GLM approach fails to account for confounding effects. Though GLM may have reasonable control over the false discoveries, its power and precision are highly affected by the confounders. On the contrary, GCATE obtain valid Type-I error and FDP controls and higher power and precision with small sample sizes because of proper distributional modeling. The result of GCATE is slightly conservative because of model misspecification and zero inflation induced by the Splatter simulator, which could bias the estimates of coefficients $\boldsymbol{B}$ towards zero. Additionally, the NB distribution involves the additional challenge of estimating the overdispersion parameters.

2.6 Lupus data example

2.6.1 The dataset

Systemic lupus erythematosus (SLE) is an autoimmune disease predominantly affecting women and individuals of Asian, African, and Hispanic descent. Perez et al. [137] developed multiplexed single-cell RNA sequencing (mux-seq) to capture the complexity of immune cell populations and systematically profile the composition and transcriptional states of immune cells in a large multiethnic cohort. The dataset contains 1.2 million peripheral blood mononuclear cells from 8 major cell types and 261 individuals, including 162 SLE cases and 99 healthy controls of either Asian or European ancestry. The cell-type-specific DE analysis aims to provide insights into the diagnosis and treatment of SLE.

To remove the genes with small variations, we use the Python package `scanpy` [178] to preprocess the single-cell data and select the top 2,000 highly variable genes (HVGs) within each cell type. For each cell type, we aggregate expression across cells from the same subject to obtain gene-level pseudo-bulk counts and then remove genes expressed in less than 10 subjects. The basic information of the preprocessed datasets is provided in Appendix A.7.3. For each subject, the recorded variables are SLE status (condition), the logarithm of the library size, sex, population, and processing cohorts (4 levels). The latter 3 variables, which account for $r = 5$ degrees of freedom, are considered to be the measured confounders.

2.6.2 Confounder adjustment

For each cell type, we compare four approaches GLM, GCATE, CATE, and CATE-mad. The first two approaches are based on the NB GLM model, while the latter two are designed for linear models. The last method uses an estimated empirical null [175] based on median absolute deviation (MAD). For each method, we consider two variants, using a "subset" of covariates and a "full" set of covariates, without and with measured confounders included, respectively. Only 5 cell types

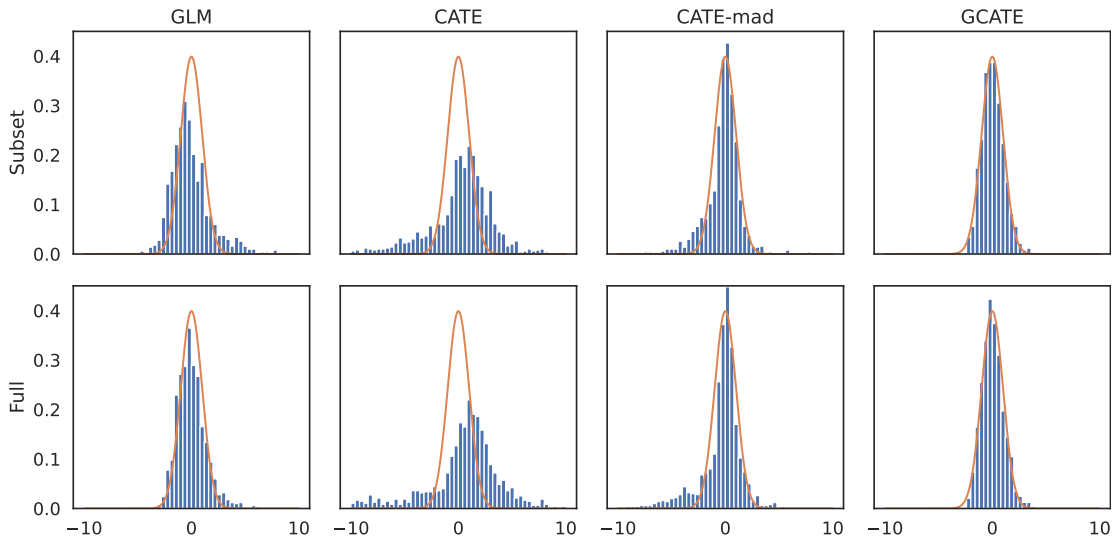

Figure 2.6: Results on the lupus datasets. Histograms of lupus z-statistics of different methods on T4 cell type. The first row uses only a subset of the covariates, while the second row uses the full set of covariates for all the methods. The orange curves represent the standard normal density.

(T4, cM, B cell, T8, NK) contain more than 50,000 single cells and have sufficient power to obtain significant findings using the GLM-full approach, so we restrict our comparisons to those types. In particular, we display our results for the largest T4 cell type in this section, and similar results for other cell types are included in Appendix A.7. To estimate the number of latent factors r, we analyze the JIC values according to Remark 1. For a subset of covariates, as shown in Figure A.75, the scree plot reveals a diminishing negative log-likelihood with increasing r, which plateaus for $r = 4$ to $r = 7$, and the decrement becomes marginal beyond $r = 7$. Consequently, we recommend selecting $r = 7$ for GCATE-subset analysis, and similarly, $r = 2$ for GCATE-full analysis. We also conduct the sensitivity analysis for the number of latent factors in Appendix A.7.4.

Our first analysis is to treat GLM-full as GLM-oracle and inspect the performance of all four methods without measured confounders included. The majority of the test statistics obtained for GLM-full are well approximated by a standard normal distribution, which suggests that the experiment conducted by [137] was well controlled, and the impact of unmeasured confounders was negligible (Figure 2.6). However, when we excluded the measured confounders, the GLM-subset statistics were poorly calibrated, indicating that controlling for these variables is essential to proper analysis, either directly or indirectly. The CATE statistics are even more poorly calibrated than GLM-subset, suggesting that these sparse data cannot be modeled using a linear model, though restricting the test to the top 250 HVGs yields test statistics closer to the expected distribution (Figure A.77). With the empirical null adjustment, CATE-mad performed somewhat better, but this adaptation is insufficient, suggesting that CATE cannot remove the confounding effects when the data are unsuitable for a linear model. Finally, the performance of GCATE is ideal: the majority of the statistics are well approximated by the standard normal, and a few signals can be captured on the right tail. Similar results were obtained for each of the 5 biggest cell types, as shown in Figure A.78.

For comparison, we label genes based on the GLM-full analysis with FDR control at cutoff 0.2 as "true positives", resulting in 72 significant genes for the T4 cell type. With FDR control at cutoff 0.2, 15 of the 16 GCATE's statistics overlap with the true positives, indicating that the test

loses power when the impact of confounders has to be removed using factor analysis. Still, the test appears to control the error rate. To illustrate the performance of the 4 competing analysis methods across all 5 large cell types, we calculate the precision and specificity using 0.2 as a cutoff for false discovery rate control. As shown in Figure A.79, only GCATE achieves uniformly high precision and specificity.

To compare different methods in the biological significance of the discoveries, we conduct gene ontology over-representation analysis to identify the related biological processes. As shown in Figure A.710, both GLM-full and GCATE-subset discover genes that are pertinent to the immune-response-related pathways, which also appear in prior studies on lupus [137, Fig. 3]. On the other hand, though hundreds of significant genes are claimed by CATE-mad, they are not associated with meaningful biological pathways. The results indicate that GCATE identifies scientifically more relevant genes than CATE under unmeasured confounders.

Our second analysis is to compare the four methods when all the measured covariates are included. As shown in the second row of Figure 2.6, we observe similar performance for each of the three methods (CATE, CATE-mad, and GCATE) remains similar whether partial or all covariates are included. In particular, we see that the test results of CATE and CATE-mad get more anti-conservative. On the other hand, as shown in Figure A.711, the results of GCATE are consistent and more powerful with added covariates, although they exhibit lower power than GLM. This is expected as, in general, the estimated latent factors may remove some signals for confounder adjustment methods. Furthermore, the GO analysis of the biological processes given in Figure A.712 aligns closely with the results in the first analysis using a subset of the covariates, suggesting the biological relevance of the findings from GCATE. Overall, GCATE demonstrates robust performance and consistency across various levels of confounding.

Chapter 3

Causal Inference for Genomic Data with Multiple Heterogeneous Outcomes

Material in this chapter first appeared as Du et al. [47, 49].

3.1 Introduction

In observational studies, causal inference on multiple outcomes is increasingly prevalent in scientific discoveries [71]. Recent advances in high-throughput techniques have enabled the collection of large-scale repeated measurements across multiple subjects in various domains. However, subject-level outcomes, such as averages or inter-correlations of measurements within each subject, are often unobserved. Instead, researchers rely on repeated measurements to construct estimates of these outcomes, referred to as *derived outcomes* (Figure 3.1). For example, advancements in single-cell RNA sequencing (scRNA-seq) techniques [50] have enabled large-scale repeated gene expression measurements across multiple cells for each individual. These measurements allow researchers to construct derived outcomes (e.g., the sample average of gene expressions for an individual) as proxies for subject-level outcomes (e.g., the true average gene expression for that individual), facilitating individual-level comparisons [185]. The goal of subject-level causal inference is to determine which unobserved outcomes are causally affected by treatment by comparing derived outcomes between treatment and control groups. However, challenges such as unobservability of outcomes, subject heterogeneity [139], and non-identical outcome distributions limit the reliability of existing causal inference methods.

One major challenge of subject-level causal inference on scRNA-seq data is the unobservability of the outcomes. Individual gene expression levels of subjects are not directly measurable; instead, repeated measurements from heterogeneous cells are aggregated to serve as proxies. Additionally, variability among subjects may arise from latent states unique to each individual that influence gene expression patterns but remain hidden from direct observation [48]. Consequently, derived outcomes often violate the classic assumption of being independent and identically distributed due to biological processes, experimental conditions, and inherent cellular heterogeneity. To analyze subject-level brain functional connectivities, prior work by Qiu et al. [139] attempts to estimate average treatment effects (ATEs) using inverse probability weighting (IPW) estimators [70, 167]. However, their approach relies on accurate propensity score modeling and assumes

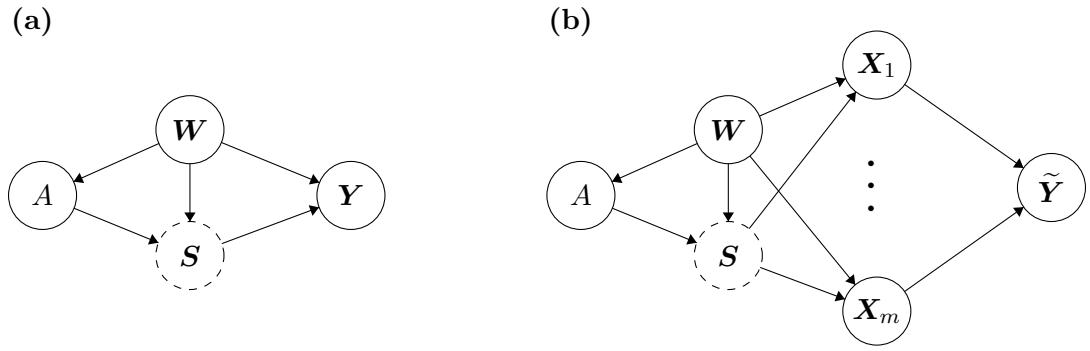

Figure 3.1: The causal diagram for the causal inference problems studied in this paper. **(a) Multiple outcomes.** For a cell, its gene expression $\boldsymbol{Y} \in \mathbb{R}^p$ is causally affected by the treatment $A \in \mathbb{R}$, the latent state $\boldsymbol{S} \in \mathbb{R}^\ell$ and covariate $\boldsymbol{W} \in \mathbb{R}^q$ such as batch effects. **(b) Multiple derived outcomes.** In the subject-level studies, a subject's overall gene expression $\boldsymbol{Y}$ is not directly observed. Instead, repeated measurements of gene expressions $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m \in \mathbb{R}^d$ of m cells from the subject provides a proxy $\widetilde{\boldsymbol{Y}}$ for $\boldsymbol{Y}$. See Section 3.3 for formal definitions. Note that the treatment effect of A on $\boldsymbol{Y}$ (or $\widetilde{\boldsymbol{Y}}$) is mediated by the latent state $\boldsymbol{S}$ even conditioned on the covariate $\boldsymbol{W}$. When conditioned on $\boldsymbol{W}$ and A, the outcomes $Y_1, \ldots, Y_p$ within the same subject are still not independent and identically distributed.

outcome homogeneity, which may not hold for genomic data.

A second challenge arises from the heterogeneity of gene expression data, which often exhibit variable scaling and right-skewed distributions, complicating comparisons across outcomes. This heterogeneity challenges the common use of ATEs since relying solely on mean differences in counterfactual distributions can lead to misleading conclusions. To estimate the treatment effects, one can rely on propensity score modeling or outcome modeling. For instance, one common strategy in scRNA-seq analyses is to model outcomes directly using parameter models such as Poisson or Negative Binomial models [150], or zero-inflated models [77]. While using either IPW or regression estimators may seem intuitive, they are both sensitive to model specification.

Doubly robust (DR) estimators [143, 152] offer a promising solution to mitigate sensitivity to model specification by combining IPW and outcome modeling. DR estimators are consistent as long as either of the two nuisance estimators is consistent, and $\sqrt{n}$-consistent whenever the nuisance estimators converge at only $n^{-1/4}$ rates (or more generally if the product of their errors is of the order $n^{-1/2}$). Additionally, if both the nuisance models are correctly specified, in the sense that the product of their errors is smaller order than $n^{-1/2}$, the DR estimators achieve the semiparametric efficiency bound for the unrestricted model, allowing the regression and propensity score functions to be estimated flexibly at slower than $n^{-1/2}$ rates in a wide variety of settings [91, 167]. Recent work introduces a structure-agnostic framework for functional estimation, demonstrating that DR estimators are optimal for estimating ATEs when only black-box estimators of the outcome model and propensity score are available [10, 78]. These results suggest that DR estimators cannot be improved upon without making additional structural assumptions. Given these advantages, exploring doubly robust estimation in settings with multiple heterogeneous outcomes is crucial, as it mitigates model misspecification and enables reliable statistical testing when nonparametric methods are used for outcome and propensity score estimation.

The unobservability of subject-level outcomes, heterogeneity in gene expression distributions, and sensitivity to model specification collectively challenge traditional causal inference methods in

scRNA-seq studies. To address these issues, we propose a semiparametric inference framework to handle multiple derived outcomes effectively. Specifically, (i) we define causal estimands that capture meaningful counterfactual differences across multiple outcomes and establish identification conditions under hierarchical models where outcomes of interest are unobserved (Figure 3.1(b)); and (ii) we develop robust and efficient estimators tailored for these estimands. Additionally, we extend multiple-testing procedures to control statistical errors during simultaneous inference on high-dimensional derived outcomes. Together, these methodological advancements provide a unified approach incorporating doubly robust estimation to handle multiple derived outcomes effectively.

Focusing on multiple derived outcomes, we first establish generic results on semiparametric inference with doubly robust estimators. It also encompasses the usual setting of multiple outcomes when the response of each unit is available. By utilizing finite-sample maximal inequality for finite maximums, we obtain interpretable conditions of uniform estimation error control for the empirical process terms and the asymptotic variances. We derive the uniform convergence rates, in terms of only finitely many moments of the influence functions' envelope and the maximum second moments of the individual estimation errors, allowing for the number of outcomes p to be potentially exponentially larger than the sample size n.

To address the challenges of outcome heterogeneity in single-cell data, we further specialize our analysis to standardized average treatment effects for comparing treatment effects across different outcomes on a common scale and quantile treatment effects for robustness against outliers. This demonstrates how generic semiparametric inferential results for DR estimators derived from von Mises expansions and estimating equations can be applied in high-dimensional settings. Furthermore, we adapt Gaussian approximation results from Chernozhukov et al. [30] to DR estimators and implement a step-down procedure to control false discovery (exceedance) rates with guaranteed power [58].

Our exploration includes two real-world application scenarios of the proposed causal inference methods. (1) *Single-cell CRISPR perturbation analysis*: Gene expressions from single cells are compared between perturbation and control groups in CRISPR experiments to identify target genes of individual perturbations and analyze the effects of perturbations [37], as shown in Figure 3.1(a). (2) *Individual level differential expression analysis*: Aggregated gene expressions from individual subjects under two conditions (case and control) are analyzed to identify genes intrinsically affected by these conditions across subjects, corresponding to Figure 3.1(b). By applying our methods to these datasets, we demonstrate their practical utility while highlighting the strengths and limitations of different causal estimands. These findings emphasize the importance of suitable causal inference techniques for the accurate interpretation of genomic data.

Organization. In Section 3.2, we review and extend the classic semiparametric inference framework to the setting of multiple outcomes. In Section 3.3, we set up the problem of interest and discuss the identification conditions for the causal estimands. In Section 3.4, we analyze two DR estimators for standardized and quantile treatment effects and study their statistical properties. In Section 3.5, we study the statistical error of simultaneous inference and propose a multiple testing procedure for controlling the false discovery rate. In Section 3.6, we conduct simulations to validate the proposed simultaneous causal inference method. A detailed review of related work is provided in Appendix B.1.

Notation. Throughout our exposition, we will use the following notational conventions. We denote scalars in non-bold lower or upper case (e.g., X), vectors in bold upper case (e.g., $\boldsymbol{X}$), and matrices in non-italic bold upper case (e.g., $\mathbf{X}$). For $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and

$a \wedge b = \min\{a, b\}$. For any random variable X, its L_q norm is defined by $\|X\|_{L_q} = (\mathbb{E}[|X|^q])^{1/q}$ for $q = 1, \ldots, \infty$. For $p \in \mathbb{N}$, $[p] := \{1, \ldots, p\}$. For a set $\mathcal{A}$, let $|\mathcal{A}|$ be its cardinality. For (potentially random) measurable functions f, we denote expectations with respect to Z alone by $\mathbb{P}f(Z) = \int f \, \mathrm{d}\mathbb{P}$, and with respect to both Z and the observations where f is fitted on by $\mathbb{E}[f(Z)]$. The empirical expectation is denoted by $\mathbb{P}_n f(Z) = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ for i.i.d. samples $Z_1, \ldots, Z_n$ of Z. Similarly, the population and empirical variances are denoted by $\mathbb{V}$ and $\mathbb{V}_n$, respectively. We write the (conditional) L_p norm of f as $\|f\|_{L_p} = \left[\int f(z)^p \, \mathrm{d}\mathbb{P}(z)\right]^{1/p}$ for $p \geq 1$. We use "o" and "$\mathcal{O}$" to denote the little-o and big-O notations and let "$o_{\mathbb{P}}$" and "$\mathcal{O}_{\mathbb{P}}$" be their probabilistic counterparts. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ or $b_n \gg a_n$ if $a_n = o(b_n)$; $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if $a_n = \mathcal{O}(b_n)$; and $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. All the constants c, c_1, c_2 and C, C_1, C_2 may vary from line to line. Convergence in distribution and probability are denoted by "$\xrightarrow{\mathrm{d}}$" and "$\xrightarrow{\mathrm{p}}$".

3.2 Semiparametric inference with multiple outcomes

Prior to delving into our main topic, this section takes a brief excursion into the formulation of semiparametric inference within the context of multiple outcomes, laying the foundation for addressing our specific problem in subsequent sections.

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be observations of i.i.d. samples from a population $\mathcal{P}$. In the presence of multiple outcomes, we are interested in estimating p target estimands $\tau_j : \mathcal{P} \mapsto \mathbb{R}$ for $j = 1, \ldots, p$. Suppose τ_j admits a von Mises expansion; that is, there exists an influence function $\varphi_j(z; \mathbb{P})$ with $\int \varphi_j(z; \mathbb{P}) \, \mathrm{d}\mathbb{P}(z) = 0$ and $\int \varphi_j(z; \mathbb{P})^2 \, \mathrm{d}\mathbb{P}(z) < \infty$, such that

$$\tau_j(\overline{\mathbb{P}}) - \tau_j(\mathbb{P}) = \int \varphi_j(z; \overline{\mathbb{P}}) \, \mathrm{d}(\overline{\mathbb{P}} - \mathbb{P}) + T_{\mathrm{R},j},$$

where $T_{\mathrm{R},j}$ is a second-order remainder term (which means it only depends on products or squares of differences between $\overline{\mathbb{P}}$ and $\mathbb{P}$). The influence function quantifies the effect of an infinitesimal contamination at the point z on the estimate, standardized by the mass of the contamination. Its historical development and definition under diverse sets of regularity conditions can be found at Hampel et al. [63, Section 2.1]. The above expansion suggests a one-step estimator that corrects the bias of the plug-in estimator $\tau_j(\widehat{\mathbb{P}})$:

$$\widehat{\tau}_j(\mathbb{P}) := \tau_j(\widehat{\mathbb{P}}) + \mathbb{P}_n\{\varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}})\}, \tag{3.2.1}$$

where $\widehat{\mathbb{P}}$ is an estimate of $\mathbb{P}$. For many estimands, such as ATE and expected conditional covariance, the one-step estimator is also a DR estimator. Although for certain estimands like expected density, the standard one-step estimator is not doubly robust, it still has nuisance errors that consist of a second-order term [87]. Then, the one-step estimator $\widehat{\tau}_j$ of the jth target estimand τ_j admits a three-term decomposition of the estimation error [84, Equation (10)][1]:

$$\widehat{\tau}_j(\mathbb{P}) - \tau_j(\mathbb{P}) = \underbrace{(\mathbb{P}_n - \mathbb{P})\{\varphi_j(\boldsymbol{Z}; \mathbb{P})\}}_{T_{\mathrm{S},j}} + \underbrace{(\mathbb{P}_n - \mathbb{P})\{\varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \varphi_j(\boldsymbol{Z}; \mathbb{P})\}}_{T_{\mathrm{E},j}} + T_{\mathrm{R},j}. \tag{3.2.2}$$

In the above decomposition, the first term after $\sqrt{n}$-scaling has an asymptotic normal distribution by the central limit theorem. That is, $\sqrt{n}\, T_{\mathrm{S},j} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathbb{V}[\varphi_j(\boldsymbol{Z}; \mathbb{P})])$. The higher-order term $T_{\mathrm{R},j}$

[1] For certain estimands, such as average treatment effects and expected conditional covariance functionals, it is usually more convenient to use the uncentered influence functions $\phi_j(\boldsymbol{Z}; \mathbb{P}) := \varphi_j(\boldsymbol{Z}; \mathbb{P}) + \tau_j(\mathbb{P})$ in the decomposition.

is usually negligible and has an order of $o_{\mathbb{P}}(n^{-1/2})$ under certain conditions. On the other hand, the empirical process term $T_{\mathrm{E},j}$ will be asymptotically negligible (i.e., $o_{\mathbb{P}}(n^{-1/2})$) under Donsker assumption [169] or sample splitting [32, 86], because it is a sample average with a shrinking variance. In our problem setting with an increasing number of outcomes, uniform control over all the empirical process terms $T_{\mathrm{E},j}$ for $j = 1, \ldots, p$ is desired to facilitate the construction of simultaneous confidence intervals. Below is an extension of Lemma 2 from Kennedy et al. [86, Appendix B] to the setting of multiple outcomes.

Lemma 8 (Uniform control of the empirical process terms)**.** Let $\mathbb{P}_n$ denote the emprical measure over $\mathcal{D}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) \in \mathcal{Z}^n$, and let $g_j : \mathcal{Z} \to \mathbb{R}$ be a (possibly random) function independent of $\mathcal{D}_n$ for $j = 1, \ldots, p$ with $p \geq 2$. Let $G(\cdot) := \max_{1 \leq j \leq p} |g_j(\cdot)|$ denote the envelope of $g_1, \ldots, g_p$. If $\max_{1 \leq j \leq p} \|g_j\|_{L_2} < \infty$ and $\|G\|_{L_q} < \infty$ for some $q \in \mathbb{N} \cup \{\infty\}$, then the following statements hold:

$$\mathbb{E}\left[\max_{j=1,\ldots,p} |(\mathbb{P}_n - \mathbb{P})g_j| \,\middle|\, \{g_j\}_{j=1}^p\right] \lesssim \left(\frac{\log p}{n}\right)^{1/2} \max_{1 \leq j \leq p} \|g_j\|_{L_2} + \left(\frac{\log p}{n}\right)^{1-1/q} \|G\|_{L_q}.$$

The proof of Lemma 8 utilizes a finite-sample maximal inequality established in Kuchibhotla and Patra [90] for high-dimensional estimation problems. When specified to particular target estimands, Lemma 8 suggests that $T_{\mathrm{E},j}$ in (3.2.2) can be uniformly controlled over $j = 1, \ldots, p$, if one can derive the uniform L_2-norm bound and the L_q-norm bound of the envelope for the estimation error of the influence functions $g_j = \varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \varphi_j(\boldsymbol{Z}; \mathbb{P})$. In particular, if g_j's are bounded, and $\log p \cdot (\max_{1 \leq j \leq p} \|g_j\|_{L_2}^2 \vee n^{-1/2}) = o(1)$, then we have that $\max_{1 \leq j \leq p} T_{\mathrm{E},j} = o_{\mathbb{P}}(n^{-1/2})$, which is negliable after scaled by $\sqrt{n}$. This allows the number of outcomes p to be potentially exponentially larger than the number of samples n. It is important to note that similar bounds on the empirical process term can still be derived even when the nuisance functions are not trained on an independent sample, provided certain complexity measures of the function class $\mathcal{F}_j$ that g_j belongs to are properly bounded; see Remark 14 in Appendix B.2.1 for more details.

Remark 5 (One-step estimator from estimating equations)**.** Above, we construct the one-step estimator based on the influence function φ_j of τ_j from von Mises expansion. One can also construct efficient estimators of pathwise differentiable functionals through estimating equations, which is related to the quantile estimand as we will discuss in Section 3.4.2.

When $T_{\mathrm{E},j} + T_{\mathrm{R},j} = o_{\mathbb{P}}(n^{-1/2})$, by central limit theorem we have that $\sqrt{n}(\widehat{\tau}_j(\mathbb{P}) - \tau_j(\mathbb{P})) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_j^2)$ where $\sigma_j^2 = \mathbb{V}[\varphi_j(\boldsymbol{Z}; \mathbb{P})]$. To construct confidence intervals, one can use the sample variance $\widehat{\sigma}_j^2 = \mathbb{V}_n[\varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}})]$ to consistently estimate the asymptotic variance. To derive the properties of test statistics and confidence intervals in high dimensions when $p \gg n$, it is necessary to establish strong control on the uniform convergence rate of the variance estimates over $j = 1, \ldots, p$; see, for example, Qiu et al. [139, Proposition 2] and Chernozhukov et al. [30, Comment 2.2]. In this regard, the following lemma provides general conditions for bounding the uniform estimation error.

Lemma 9 (Uniform control of the variance estimates)**.** Denote $\varphi_j = \varphi_j(\boldsymbol{Z}; \mathbb{P})$, $\widehat{\varphi}_j = \varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}})$, $\Phi = \max_{1 \leq j \leq p} |\widehat{\varphi}_j - \varphi_j|$, and $\Psi = \max_{1 \leq j \leq p} |\varphi_j|$. Suppose the following conditions hold:
(1) Envelope: $\max_{1 \leq j \leq p} |\widehat{\varphi}_j + \varphi_j| \lesssim 1$ and $\|\Psi\|_{L_q} + \max_{k=1,2} \|\Phi^k\|_{L_q} \lesssim r_{1n}$ for some $q > 1$,
(2) Estimation error: $\max_{1 \leq j \leq p} \|\widehat{\varphi}_j - \varphi_j\|_{L_2} \lesssim r_{2n}$, $\max_{1 \leq j \leq p} |\mathbb{P}[\widehat{\varphi}_j - \varphi_j]| \lesssim r_{3n}$,
with probability tending to one. Then, it holds that

$$\max_{1 \leq j \leq p} |\widehat{\sigma}_j^2 - \sigma_j^2| \lesssim \mathcal{O}_{\mathbb{P}}\left(\left(\frac{\log p}{n}\right)^{1-1/q} r_{1n} + \left(\frac{\log p}{n}\right)^{1/2} r_{2n} + r_{3n}\right).$$

Note that r_{1n} and r_{2n} are allowed to potentially diverge. We will utilize Lemma 9 for the purpose of multiple testing, as demonstrated in Section 3.5. Typically, to accommodate an exponentially large number of outcomes p relative to n while maintaining valid statistical inference, it suffices to ensure that variance estimates are consistent at any polynomial rate of the sample size n; specifically, $\max_{1\le j\le p} |\widehat{\sigma}_j^2 - \sigma_j^2| = o_{\mathbb{P}}(n^{-\alpha})$ for some constant $\alpha > 0$.

3.3 Subject-level causal inference with multiple outcomes

We consider an increasingly popular study design where scRNA-seq data are collected from multiple individuals and the question of interest is to find genes that are causally differentially expressed between two groups of individuals, based on repeated single-cell measurements.

3.3.1 Causal inference with multiple derived outcomes

Suppose a subject can be either in the case or control group, indicated by a binary random variable $A \in \{0, 1\}$ and we sequence the expressions $\mathbf{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m]^\top \in \mathbb{R}^{m\times d}$ of d genes in m cells[2], along with subject-level covariates $\boldsymbol{W} \in \mathbb{R}^q$. Let $\mathbf{X}(a)$ denote the *potential response* of gene expressions. To characterize the complex biological processes, suppose $\boldsymbol{S}(a) \in \mathbb{R}^\ell$ is the latent potential state after receiving treatment $A = a$, which fully captures the effects of the treatment on the individual. We assume $\boldsymbol{X}_1(a), \ldots, \boldsymbol{X}_m(a)$ are conditionally independent and identically distributed given $\boldsymbol{S}(a)$ and $\boldsymbol{W}$[3]. Marginally, however, they can be highly dependent because of repeated measurements from the same individual. As an example of genomics data, $\boldsymbol{S}$ can be the chromatin accessibility that governs the translation and expression of genes, while $\boldsymbol{X}_m$ is the resulting expression level of those genes.

Suppose the collection of treatment assignment, covariates, subject level parameters, and potential responses $(A, \boldsymbol{W}, \boldsymbol{S}(0), \boldsymbol{S}(1), \mathbf{X}(0), \mathbf{X}(1))$ is from some super-population $\mathcal{P}$. We require the consistency assumption on the observed response.

Assumption 5 (Consistency)**.** The observed response is given by $\mathbf{X} = A\mathbf{X}(1) + (1 - A)\mathbf{X}(0)$.

When comparing gene expressions between two groups of individuals, the p-dimensional subject-level parameter of interest, is the *potential outcome* $\boldsymbol{Y}(a) \in \mathbb{R}^p$, a functional that maps the conditional distribution of $\boldsymbol{X}_1(a)$ given $\boldsymbol{S}(a)$ and $\boldsymbol{W}$ to $\mathbb{R}^p$:

$$\boldsymbol{Y}(a) = \mathbb{E}\left[f(\boldsymbol{X}_1(a)) \,|\, \boldsymbol{S}(a), \boldsymbol{W}\right], \tag{3.3.1}$$

for some prespecified function $f : \mathbb{R}^d \to \mathbb{R}^p$. The choice of f should align with the user's research goals and the specific aspects of the data they intend to capture. For example, when f is the identity map and $p = d$, the potential outcome $\boldsymbol{Y}(a)$ represents the conditional mean; when $f(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{X}^\top - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top \mid \boldsymbol{S}(a), \boldsymbol{W}]$ and $p = d$, it represents the intrasubject covariance matrix. When considering conditional means among the potential responses $X_{1j}(a)$'s, it can also be nodewise regression coefficients as considered in Qiu et al. [139]. Intuitively, $\boldsymbol{Y}(a)$ is an individual / within-group characteristic that depends on the conditional distribution of $\boldsymbol{X}_1(a)$ given $\boldsymbol{S}(a), \boldsymbol{W}$. From Assumption 5, we also have $\boldsymbol{Y} = A\boldsymbol{Y}(1) + (1 - A)\boldsymbol{Y}(0)$. Compared to the classical causal inference setting, the subject-level outcome $\boldsymbol{Y}$ is not observed for each subject,

[2]For notational simplicity, we treat the number of cell m as fixed across subjects, though the method also applies when the number of cell m_i varies for subjects $i = 1, \ldots, n$.

[3]Technically, the potential response can be denoted as $\boldsymbol{X}_m(\boldsymbol{S}(a))$; however, because $\boldsymbol{S}(a)$ is unobservable and the variable to intervene is A, we use a simplified notation $\boldsymbol{X}_m(a)$ to denote the potential response.

while only the repeated measurements of gene expressions $\mathbf{X}$ from multiple cells are available and can be used to construct a derived outcome $\widetilde{\boldsymbol{Y}}$. For a given f, we consider a statistic $\widetilde{\boldsymbol{Y}} := g(\mathbf{X})$ for some function $g : (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m) \mapsto \widetilde{\boldsymbol{Y}}$. There can be different choices of g to estimate $\boldsymbol{Y}(a)$ by $\widetilde{\boldsymbol{Y}}(a)$. For instance, if f is a linear function for the potential outcome $\boldsymbol{Y}(a)$ in (3.3.1), then g can be a simple sample average as a natural choice of the derived outcome; alternatively, g can also be the median-of-means estimator as the derived outcomes.

Under the derived outcomes framework, Qiu et al. [139] studied the IPW estimator for ATE:

$$\tau_j^{\mathrm{ATE}} = \mathbb{E}[Y_j(1) - Y_j(0)], \qquad j = 1, \ldots, p. \tag{3.3.2}$$

By focusing on the expected potential outcomes, we next describe the identification condition and semiparametric inferential results based on derived outcomes $\widetilde{\boldsymbol{Y}}$.

Identification. We require two extra classic causal assumptions for observational studies.

Assumption 6 (Positivity)**.** The propensity score $\pi_a(\boldsymbol{W}) := \mathbb{P}(A = a \mid \boldsymbol{W}) \in (0, 1)$.

Assumption 7 (No unmeasured confounders)**.** $A \perp\!\!\!\perp \mathbf{X}(a) \mid \boldsymbol{W}$, for all $a \in \{0, 1\}$.

The above assumptions on the propensity score and the potential responses are standard for observational studies in the causal inference literature [71, 84]. Assumption 6 suggests that both treated and control units of each subject can be found for any value of the covariate with a positive probability. Assumption 7 ensures that the treatment assignment is fully determined by the observed covariate $\boldsymbol{W}$. These assumptions are required to estimate functionals of $\mathbf{X}(a)$ with observed variables $(A, \boldsymbol{W}, \mathbf{X})$.

Let $\boldsymbol{Z} = (A, \boldsymbol{W}, \mathbf{X}, \boldsymbol{Y})$ denote the tuple of observed random variables and unobserved outcomes $\boldsymbol{Y}$. Because the outcome $\boldsymbol{Y}$ is not observed for each subject, we are interested in constructing a proxy $\widetilde{\boldsymbol{Y}} = g(\mathbf{X})$ of $\boldsymbol{Y}$ from repeated measurements $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ from the same subject. Analogously, we denote $\widetilde{\boldsymbol{Y}}(a) := g(\mathbf{X}(a))$ for the potential outcomes and quantify the bias as $\boldsymbol{\Delta}_m(a) := \mathbb{E}[\widetilde{\boldsymbol{Y}}(a) \mid \boldsymbol{W}, \boldsymbol{S}(a)] - \boldsymbol{Y}(a)$. Below, we introduce a notion of asymptotic unbiased estimate in Definition 1, where the expected bias is negligible uniformly over multiple outcomes.

Definition 1 (Asymptotic unbiased estimate)**.** For $a \in \{0, 1\}$, the derived outcome $\widetilde{\boldsymbol{Y}}(a)$ is asymptotic unbiased to $\boldsymbol{Y}(a)$ if the bias tends to zero: $\max_{1 \le j \le p} |\mathbb{E}[\Delta_{mj}(a)]| = o(1)$ as $m \to \infty$.

Note that when $\widetilde{\boldsymbol{Y}}(a)$ is marginally unbiased, i.e., $\mathbb{E}[\boldsymbol{Y}(a)] = \mathbb{E}[\widetilde{\boldsymbol{Y}}(a)]$, it also implies that $\widetilde{\boldsymbol{Y}}(a)$ is asymptotically unbiased to $\boldsymbol{Y}(a)$. Therefore, our framework also includes the common setting where all the outcomes $\boldsymbol{Y}(a) = \widetilde{\boldsymbol{Y}}(a) = \boldsymbol{X}_1(a)$ (with $m = 1$) are observed. When $\widetilde{\boldsymbol{Y}}(a)$ is an asymptotic unbiased estimate of $\boldsymbol{Y}(a)$, Lemma 1 from Qiu et al. [139] suggests that the counterfactual of unobserved outcomes can be identified asymptotically, as detailed in Proposition 10.

Proposition 10 (Identification of linear functionals)**.** Under Assumptions 5–7, if $\widetilde{\boldsymbol{Y}}(a)$ is asymptotically unbiased to $\boldsymbol{Y}(a)$, then $\mathbb{E}[\boldsymbol{Y}(a)]$ can be identified by $\mathbb{E}[\mathbb{E}[\widetilde{\boldsymbol{Y}} \mid \boldsymbol{W}, A = a]]$ as $m \to \infty$.

Semiparametric inference. When the target causal estimands are the expectation of the potential outcomes $\tau_j = \mathbb{E}[Y_j(a)]$ for $j = 1, \ldots, p$, one can adopt results from Section 3.2 to establish the asymptotic normality of certain estimators under proper assumptions on the convergence rate of the nuisance function estimates. However, because $\boldsymbol{Y}$ is unobservable, we are not able to directly estimate its influence function and hence its influence-function-based one-step estimator (3.2.1). Instead, we can rely on the influence function of $\widetilde{\tau}_j = \mathbb{E}[\widetilde{Y}_j(a)]$: $\widetilde{\varphi}_j(\boldsymbol{Z}; \mathbb{P}) = \mathbb{1}\{A = a\}\pi_a(\boldsymbol{W})^{-1}(\widetilde{Y}_j - \mu_{aj}(\boldsymbol{W})) + \mu_{aj}(\boldsymbol{W}) - \widetilde{\tau}_j(\mathbb{P})$, where $\pi_a(\boldsymbol{W}) = \mathbb{P}(A = a \mid \boldsymbol{W})$

and $\mu_{aj}(\boldsymbol{W}) = \mathbb{E}[\widetilde{Y}_j \mid A = a, \boldsymbol{W}]$ for $j = 1, \ldots, p$, are the propensity score and regression functions, respectively. This, in turn, yields an analog of the one-step estimator (3.2.1):

$$\widehat{\tau}_j(\mathbb{P}) := \widetilde{\tau}_j(\widehat{\mathbb{P}}) + \mathbb{P}_n\{\widetilde{\varphi}(\boldsymbol{Z}; \widehat{\mathbb{P}})\},$$

which further implies the decomposition of the estimation error for the causal estimand τ_j:

$$\widehat{\tau}_j(\mathbb{P}) - \tau_j(\mathbb{P}) = T_{\mathrm{S},j} + T_{\mathrm{E},j} + T_{\mathrm{R},j} + \mathbb{E}[\Delta_{mj}], \tag{3.3.3}$$

where the sample average term $T_{\mathrm{S},j}$, the empirical process term $T_{\mathrm{E},j}$ and the reminder term $T_{\mathrm{R},j}$ are as in (3.2.2) with φ_j replaced by $\widetilde{\varphi}_j$. The asymptotic variance $\sigma_j^2 = \mathbb{V}[\widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}})]$ can be estimated by the empirical variance $\widehat{\sigma}_j^2 = \mathbb{V}_n[\widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}})]$ analogously. However, the application of Lemmas 8 and 9 would require the verification of conditions for the perturbed influenced function $\widetilde{\varphi}_j$ instead of φ_j. Similar ideas apply to the one-step and doubly robust estimators of other target estimands.

3.3.2 Beyond average treatment effects

For single-cell gene expressions exhibiting different scales and skew-distributed, simply comparing the average treatment effects (3.3.2) may not be reliable. One approach to improve on the naive estimand is to consider standardized average treatment effects (STE):

$$\tau_j^{\mathrm{STE}} = \frac{\mathbb{E}[Y_j(1) - Y_j(0)]}{\sqrt{\mathbb{V}[Y_j(0)]}}, \qquad j = 1, \ldots, p \tag{3.3.4}$$

which allows for consistent and comparative analysis across different scales and variances, enhancing the interpretability and comparability of treatment effects in diverse and complex datasets [85]. Another approach is to consider quantile effects (QTE):

$$\tau_j^{\mathrm{QTE}_\varrho} = Q_\varrho[Y_j(1)] - Q_\varrho[Y_j(0)], \qquad j = 1, \ldots, p, \tag{3.3.5}$$

where $Q_\varrho[U]$ denote the ϱ-quantile of random variable U. In particular, when $\varrho = 0.5$, the ϱ-quantile equals the median $Q_\varrho(U) = \mathrm{Med}(U)$, and we reveal one of the commonly used robust estimand $\tau_j^{\mathrm{QTE}} = \mathrm{Med}[Y_j(1)] - \mathrm{Med}[Y_j(0)]$ for location-shift hypotheses. QTE may be more robust and less affected by the outliers of gene expressions [23, 81].

Note that the identification condition and semiparametric inferential results in Section 3.3.1 do not apply directly to target estimands other than ATE. Therefore, efforts are required to generalize the results to include STE and QTE for multiple derived outcomes. This demonstrates the utility and validity of our semiparametric inferential framework on one-step estimators defined through the von Mises expansion and the formulation of estimating equations, respectively, as investigated next.

3.4 Doubly robust estimation

In this section, we analyze the DR estimators for STE and QTE, which exemplify the application of general theoretical results in Section 3.2 to specific target estimands.

3.4.1 Standarized average effects

Recall the standardized average treatment effects τ_j^{STE} defined in (3.3.4), for $j = 1, \ldots, p$. The following lemma provides the identified forms of STE based on observational data.

Lemma 11 (Identification of standarized average effects)**.** Under Assumptions 5–7, if $\mathbb{V}[Y_j(0)] > 0$ and $\widetilde{Y}_j(a)^k$ is asymptotically unbiased to $Y_j(a)^k$ for $k = 1, 2$ and $a = 0, 1$ such that $k + a \leq 2$, i.e., the bias of the derived outcomes $\Delta_{mkj}(a) := \mathbb{E}[\widetilde{Y}_j(a)^k \mid \boldsymbol{W}, \boldsymbol{S}(a)] - Y_j(a)^k$ satisfies that $\delta_m := \max_{k,a} \max_{1 \leq j \leq p} |\mathbb{E}[\Delta_{mkj}(a)]| = o(1)$, then as $m \to \infty$, the standardized average treatment effect τ_j^{STE} can be identified by $\tau_j^{\mathrm{STE}} = \widetilde{\tau}_j^{\mathrm{STE}} + o_{\mathbb{P}}(1)$ where

$$\widetilde{\tau}_j^{\mathrm{STE}} := \frac{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A = 1, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A = 0, \boldsymbol{W})]}{\sqrt{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j^2 \mid A = 0, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A = 0, \boldsymbol{W})]^2}}. \tag{3.4.1}$$

As suggested by Lemma 11, estimating STE requires estimating the conditional expectation of $\widetilde{Y}_j^k$ given $A = a$ and $\boldsymbol{W}$. For this purpose, we consider the DR estimator of treatment effect $\mathbb{E}[\widetilde{Y}_j(1)]$:

$$\widetilde{\phi}_{akj}(\boldsymbol{Z}; \pi_a, \boldsymbol{\mu}_a) := \frac{\mathbb{1}\{A = a\}}{\pi_a(\boldsymbol{W})}(\widetilde{Y}_j^k - \mu_{akj}(\boldsymbol{W})) + \mu_{akj}(\boldsymbol{W}),$$

where $\boldsymbol{\mu}_a : \mathbb{R}^d \to \mathbb{R}^{2 \times p}$ is the mean regression function with entry $\mu_{akj}(\boldsymbol{W}) = \mathbb{E}[\widetilde{Y}_j^k \mid \boldsymbol{W}, A = a]$ and $\pi_a(\boldsymbol{W}) = \mathbb{P}(A = a \mid \boldsymbol{W})$ is the propensity score function. By plugging in the DR estimators for individual counterfactual expectations consisting in (3.4.1), we obtain a natural estimator for the STE:

$$\widehat{\tau}_j^{\mathrm{STE}} = \frac{\mathbb{P}_n\{\widetilde{\phi}_{11j}(\boldsymbol{Z}; \widehat{\pi}_1, \widehat{\boldsymbol{\mu}}_1) - \widetilde{\phi}_{01j}(\boldsymbol{Z}; \widehat{\pi}_0, \widehat{\boldsymbol{\mu}}_0)\}}{\sqrt{\mathbb{P}_n\{\widetilde{\phi}_{02j}(\boldsymbol{Z}; \widehat{\pi}_0, \widehat{\boldsymbol{\mu}}_0)\} - \mathbb{P}_n\{\widetilde{\phi}_{01j}(\boldsymbol{Z}; \widehat{\pi}_0, \widehat{\boldsymbol{\mu}}_0)\}^2}}, \tag{3.4.2}$$

which is also the DR estimator of $\widetilde{\tau}_j^{\mathrm{STE}}$. The following theorem shows that under mild conditions, the above estimator $\widehat{\tau}_j^{\mathrm{STE}}$ is doubly robust for estimating τ_j^{STE}, with the remainder terms uniformly controlled over all outcomes.

Theorem 12 (Linear expansion of STE)**.** Under Assumptions 5–7 and the identification condition in Lemma 11, consider the one-step estimator (3.4.2), where $\mathbb{P}_n$ is the empirical measure over $\mathcal{D} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$ and $(\widehat{\pi}_a, \widehat{\boldsymbol{\mu}}_a)$ is an estimate of $(\pi_a, \boldsymbol{\mu}_a)$ for $a = 0, 1$ from samples independent of $\mathcal{D}$. Suppose the following hold for $k = 1, 2$ and $a = 0, 1$ with probability tending to one:

(1) Boundedness: There exists $c, C > 0$ and $\epsilon \in (0, 1)$ such that $\max\{|Y_j|, |\widetilde{Y}_j|\} < C$, $\max\{\|\mu_{akj}\|_{L_\infty}, \|\widehat{\mu}_{akj}\|_{L_\infty}\} < C$, $\mathbb{V}[Y_j(0)] > c$ for all $j \in [p]$, and $\pi_a, \widehat{\pi}_a \in [\epsilon, 1 - \epsilon]$.

(2) Nuisance: The rates of nuisance estimates are $\max_{j \in [p]} \|\widehat{\mu}_{akj} - \mu_{akj}\|_{L_2} = \mathcal{O}(n^{-\alpha})$ and $\|\widehat{\pi}_a - \pi_a\|_{L_2} = \mathcal{O}(n^{-\beta})$ for some $\alpha, \beta \in (0, 1/2)$ such that $\alpha + \beta > 1/2$.

Then as $m, n, p \to \infty$, it holds that $\widehat{\tau}_j^{\mathrm{STE}} - \tau_j^{\mathrm{STE}} = \mathbb{P}_n\{\widetilde{\varphi}_j^{\mathrm{STE}}\} + \varepsilon_j$, $j = 1, \ldots, p$, where the residual terms satisfy $\max_{j \in [p]} |\varepsilon_j| = \mathcal{O}_{\mathbb{P}}(n^{-(\alpha+\beta)} + \vartheta^{\mathrm{STE}}\sqrt{(\log p)/n} + (\log p)/n + \delta_m)$ with $\vartheta^{\mathrm{STE}} := n^{-(\alpha \wedge \beta)}$ and the influence function is given by

$$\widetilde{\varphi}_j^{\mathrm{STE}} = \frac{\widetilde{\phi}_{11j} - \widetilde{\phi}_{01j}}{\sqrt{\mathbb{V}[\widetilde{Y}_j(0)]}} - \widetilde{\tau}_j^{\mathrm{STE}} \left[\frac{\widetilde{\phi}_{02j} + \mathbb{E}[\widetilde{Y}_j(0)^2] - 2\mathbb{E}[\widetilde{Y}_j(0)]\widetilde{\phi}_{01j}}{2\,\mathbb{V}[\widetilde{Y}_j(0)]} \right]. \tag{3.4.3}$$

The proof of Theorem 12 requires the analysis of the linear expansions for the individual counterfactual expectations $\mathbb{E}[Y_j^k(a)]$ (see Lemma B.4.1). It then requires the application of the delta method to derive the uniform convergence rates of the residuals over multiple outcomes. For the residuals' rate, the term $n^{-(\alpha+\beta)}$ is the product of the two nuisance estimation errors, which shows the benefit of the double robustness property, while the term $\vartheta^{\mathrm{STE}}\sqrt{(\log p)/n}+(\log p)/n$ is related to the empirical process terms of individual counterfactuals by applying Lemma 8. From triangular-array central limit theorem (Lemma B.4.4), a direct consequence of Theorem 12 is the asymptotic normality of individual STE estimators, as presented in Corollary 13.

Corollary 13 (Asymptotic normality)**.** Under conditions in Theorem 12, when $(\vartheta^{\mathrm{STE}}\vee n^{-1/4})\sqrt{\log p} = o(1)$, $\delta_m = o(n^{-1/2})$ and $\sigma_j^2 := \mathbb{V}(\widetilde{\varphi}_j^{\mathrm{STE}}(\boldsymbol{Z};\pi,\boldsymbol{\mu})) \geq c$ for some constant $c>0$, it holds that,

$$\sqrt{n}(\widehat{\tau}_j^{\mathrm{STE}} - \tau_j^{\mathrm{STE}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma_j^2), \qquad j=1,\ldots,p.$$

Compared to Definition 1, Corollary 13 requires a stronger condition on the rate of the bias δ_m, which is mild. For instance, when $\boldsymbol{X}_1(a),\ldots,\boldsymbol{X}_m(a)$ are i.i.d. conditional on $(\boldsymbol{W},\boldsymbol{S}(a))$, the bias is zero, i.e., $\mathbb{E}[\Delta_{mkj}(a)]\equiv 0$; when they are weakly dependent, for example, Qiu et al. [139, Proposition S1] show that the bias is of order $o(n^{-1/2})$ under the β-mixing condition when $n^{1/2}\log p = o(m)$.

Because the influence function of STE (3.4.2) is a complicated function of all the nuisances and the observations in $\mathcal{D}$, it is hard to show usual sample variance of the estimated influence function $\widehat{\varphi}_j^{\mathrm{STE}} = \widetilde{\varphi}_j^{\mathrm{STE}}(\boldsymbol{Z};\{\widehat{\pi}_a,\widehat{\boldsymbol{\mu}}_a\}_{a\in\{0,1\}})$ provides a consistent estimate. In the following proposition, we thus rely on extra independent observations to estimate the asymptotic variance. However, one can employ the cross-fitting procedure [32] on $\mathcal{D}$ to decouple the dependency of $\widehat{\varphi}_j^{\mathrm{STE}}$ and the observations used to compute the empirical variance. This ensures that the variance estimation errors are of polynomial rates of n uniformly in p outcomes when $\log(p)/n \leq Cn^{-c}$.

Proposition 14 (Consistent variance estimates)**.** Under the same conditions in Theorem 12, let $\widehat{\varphi}_j^{\mathrm{STE}}$ be the estimated influence function (3.4.3) with $(\mathbb{E}[\widetilde{Y}_j(0)],\mathbb{E}[\widetilde{Y}_j(0)^2],\widetilde{\tau}_j^{\mathrm{STE}})$ estimated by the doubly robust estimators on $\mathcal{D}$, and $\mathbb{P}_n'$ be the empirical measure over a separate independent sample $\mathcal{D}' = \{\boldsymbol{Z}_{n+1},\ldots,\boldsymbol{Z}_{2n}\}$. Define the sample variance on $\mathcal{D}'$ as $\widehat{\sigma}_j^2 = \mathbb{V}_n'(\widehat{\varphi}_j^{\mathrm{STE}})$. It holds that $\max_{j\in[p]}|\widehat{\sigma}_j^2 - \sigma_j^2| = \mathcal{O}_{\mathbb{P}}(r_\sigma^{\mathrm{STE}})$ where $r_\sigma^{\mathrm{STE}} = \sqrt{\log p/n} + \vartheta^{\mathrm{STE}}$.

3.4.2 Quantile effects

In practice, examining quantile effects offers a robust alternative to mean-based analysis, particularly when confronted with highly variable treatment assignment probabilities and heavy-tailed outcomes. Estimating causal effects on the mean is a challenging problem in such scenarios because the signal-noise ratio is generally small. In cases where the mean is undefined but the median exists (such as the Cauchy distribution), using the median may result in more powerful tests for the location-shift hypothesis [36].

We first introduce the DR estimator for the median effect (3.3.5) when $\varrho = 0.5$, while the proposal naturally extends to other quantile levels ϱ as well. For $j\in[p]$, let θ_{aj} be the ϱ-quantile of the counterfactual response $Y_j(a)$, which solves the following equation:

$$0 = \mathbb{E}[\psi(Y_j(a),\theta)], \text{ where } \psi(y,\theta) := \mathbb{1}\{y\leq\theta\} - \varrho. \tag{3.4.4}$$

Since the potential outcome $Y_j(a)$ is not directly observed, we need to rely on the counterfactual derived outcomes $\widetilde{Y}_j(a)$ to identify the quantile of $Y_j(a)$. The following lemma summarizes the

identification results of general M-estimators for functionals of $Y_j(a)$ using the derived outcomes $\widetilde{Y}_j(a)$.

Lemma 15 (Identification of M-estimators)**.** Under Assumptions 5–7, consider the causal estimand θ_{aj} as the solution to the estimation equations:

$$M_j(\theta) = \mathbb{E}[F_j(Y_j(a), \theta)] = 0, \qquad j = 1, \ldots, p,$$

where M_j is differentiable and the magnitude of its derivative $|M_j'|$ is uniformly lower bounded around θ_{aj}: $\min_{1\le j\le p} \inf_{\theta\in\mathcal{B}(\theta_{aj},\delta)} |M_j'(\theta)| \ge c > 0$ for some constant $\delta > 0$. Suppose $\widetilde{Y}_j(a)$'s are derived outcomes such that $F_j(\widetilde{Y}_j(a), \theta)$ is asymptotically unbiased to $F_j(Y_j(a), \theta)$, i.e. as $m \to \infty$, $\Delta_{mj}(a, \theta) = \mathbb{E}[F_j(\widetilde{Y}_j(a), \theta) \mid \boldsymbol{S}(a), \boldsymbol{W}] - F_j(Y_j(a), \theta)$ satisfies that $\delta_m := \max_{j\in[p]} \sup_{\theta\in\mathcal{B}(\theta_{aj},\delta)} |\mathbb{E}[\Delta_{mj}(a, \theta)]| = o(1)$. Let $\widetilde{\theta}_{aj} \in \mathcal{B}(\theta_{aj}, \delta)$ be the solution to the estimating equation

$$\mathbb{E}[\mathbb{E}[F_j(\widetilde{Y}_j, \theta) \mid A = a, \boldsymbol{W}]] = 0,$$

then θ_{aj} can be identified by $\widetilde{\theta}_{aj}$ as $m \to \infty$.

Under conditions in Lemma 15 with $F_j(Y_j(a), \theta) = \psi(Y_j(a), \theta)$, we can focus on estimating the quantile of $\widetilde{Y}_j(a)$ to approximate the quantile of $Y_j(a)$. Specifically, consider a doubly robust expansion of the above question: $0 = \mathbb{E}[\psi(\widetilde{Y}_j, \theta)] = -\mathbb{E}[\omega_{aj}(\boldsymbol{Z}, \theta)]$, where the estimating function is given by

$$\omega_{aj}(\boldsymbol{Z}, \theta) = \frac{\mathbb{1}\{A = a\}}{\pi_a(\boldsymbol{W})}(\nu_{aj}(\boldsymbol{W}, \theta) - \psi(\widetilde{Y}_j, \theta)) - \nu_{aj}(\boldsymbol{W}, \theta).$$

Here, $\nu_{aj}(\boldsymbol{W}, \theta) = \mathbb{E}[\psi(\widetilde{Y}_j, \theta) \mid \boldsymbol{W}, A = a] = \mathbb{P}(\widetilde{Y}_j \le \theta \mid \boldsymbol{W}, A = a) - \varrho$ is the excess (conditional) cumulative distribution functions (cdfs) of $\widetilde{Y}_j(a)$, and π_a is the propensity score function as before. One may then expect to obtain an estimator of θ_{aj} by solving the empirical version of (3.4.4) for θ:

$$0 = \mathbb{P}_n[\widehat{\omega}_{aj}(\boldsymbol{Z}, \theta)], \tag{3.4.5}$$

where $\widehat{\omega}_{aj}(\boldsymbol{Z}, \theta) = \frac{\mathbb{1}\{A=a\}}{\widehat{\pi}_a(\boldsymbol{W})}(\widehat{\nu}_{aj}(\boldsymbol{W}, \theta) - \psi(\widetilde{Y}_j, \theta)) - \widehat{\nu}_{aj}(\boldsymbol{W}, \theta)$ is the estimated influence function and $(\widehat{\pi}_a, \widehat{\nu}_{aj} + \varrho)$ are the estimated propensity score and cdf functions with range $[0, 1]$.

However, directly solving (3.4.5) is not straightforward due to its non-smoothness and non-linearity in θ. A reasonable strategy to adopt instead is a one-step update approach [167, 169] using the influence function:

$$\widehat{\theta}_{aj} = \widehat{\theta}_{aj}^{\text{init}} + \frac{1}{\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})}\mathbb{P}_n[\widehat{\omega}_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})], \tag{3.4.6}$$

where $\widehat{\theta}_{aj}^{\text{init}}$ is an initial estimator of θ_{aj} and $\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})$ is the estimated density of $\widetilde{Y}_j(a)$ at $\widehat{\theta}_{aj}^{\text{init}}$.

For $a = 0, 1$, let $\boldsymbol{f}_a = (f_{aj})_{j\in[p]}$, $\widetilde{\boldsymbol{\theta}}_a = (\widetilde{\theta}_{aj})_{j\in[p]}$, and $\boldsymbol{\nu}_a = (\nu_{aj})_{j\in[p]}$ be the vectors of true density functions, the ϱ-quantiles, and the excess cdf functions cdfs of $\widetilde{Y}_j(a)$, respectively. Moreover, let $\widehat{\boldsymbol{f}}_a = (\widehat{f}_{aj})_{j\in[p]}$, $\widehat{\boldsymbol{\theta}}_a = (\widehat{\theta}_{aj})_{j\in[p]}$, and $\widehat{\boldsymbol{\nu}}_a = (\widehat{\nu}_{aj})_{j\in[p]}$ be the corresponding vectors of estimated nuisances. Based on (3.4.6), an estimator for τ_j^{QTE} is given by

$$\widehat{\tau}_j^{\text{QTE}} = \widehat{\theta}_{1j} - \widehat{\theta}_{0j}. \tag{3.4.7}$$

The following theorem provides the asymptotic normality of the one-step estimator (3.4.7).

Theorem 16 (Linear expansion of QTE)**.** Under Assumptions 5–7, suppose the identification conditions in Lemma 15 hold with $F_j = \psi$ defined in (3.4.4). Consider the one-step estimator (3.4.7) for the median treatment effect, where $\mathbb{P}_n$ is the empirical measure over $\mathcal{D} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$ and $(\widehat{\boldsymbol{\theta}}_a^{\text{init}}, \widehat{\boldsymbol{f}}_a, \widehat{\pi}_a, \widehat{\boldsymbol{\nu}}_a)$ is an estimate of $(\widetilde{\boldsymbol{\theta}}_a, \boldsymbol{f}_a, \pi_a, \boldsymbol{\nu}_a)$ from samples independent of $\mathcal{D}$ for $a = 0, 1$. Suppose the following conditions hold for $a = 0, 1$ with probability tending to one:

(1) Boundedness: The quantile θ_{aj} is in the interior of its parameter space. There exists $C, c > 0$ and $\epsilon, \delta \in (0, 1)$ such that $\max_{1\le j\le p} \max\{|Y_j|, |\widetilde{Y}_j|\} < C$, and $\pi_a, \widehat{\pi}_a \in [\epsilon, 1-\epsilon]$, and f_{aj} is uniformly bounded : $c \le f_{aj} \le C$ for all j and has a bounded derivative in a neighborhood $\mathcal{B}(\widetilde{\theta}_{aj}, \delta)$ for all $j \in [p]$: $\max_{1\le j\le p} \max_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)} |f'_{aj}(\theta)| \le C$.

(2) Initial estimation: The initial quantile and density estimators satisfy that $\max_{j\in[p]} |\widehat{\theta}_{aj}^{\text{init}} - \widetilde{\theta}_{aj}| = \mathcal{O}(n^{-\gamma})$ and $\max_{j\in[p]} |\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}}) - f_{aj}(\widetilde{\theta}_{aj})| = \mathcal{O}(n^{-\kappa})$ with $\gamma > 1/4, \kappa > 0$ such that $\gamma + \kappa > 1/2$.

(3) Nuisance: The rates of nuisance estimates satisfy $\max_{j\in[p]} \sup_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)} \|\widehat{\nu}_{aj}(\cdot,\theta) - \nu_{aj}(\cdot,\theta)\|_{L_2} = \mathcal{O}(n^{-\alpha})$ and $\|\widehat{\pi}_a - \pi_a\|_{L_2} = \mathcal{O}(n^{-\beta})$ for some $\alpha, \beta \in (0, 1/2)$ such that $\alpha + \beta > 1/2$.

Then as $m, n, p \to \infty$, it holds that $\widehat{\tau}_j^{\text{QTE}} - \tau_j^{\text{QTE}} = \mathbb{P}_n\{\widetilde{\varphi}_j^{\text{QTE}}\} + \varepsilon_j$, $j = 1, \ldots, p$, where the residual term satisfy $\max_{j\in[p]} |\varepsilon_j| = \mathcal{O}_{\mathbb{P}}(\vartheta^{\text{QTE}}\sqrt{(\log p)/n} + (\log p)/n + n^{-(\alpha+\beta)\wedge(\gamma+\kappa)\wedge(2\gamma)} + \delta_m)$ with $\vartheta^{\text{QTE}} := n^{-(\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2})}$ and the influence function is given by

$$\widetilde{\varphi}_j^{\text{QTE}}(\boldsymbol{Z}; \{\widetilde{\boldsymbol{\theta}}_a, \boldsymbol{f}_a, \pi_a, \boldsymbol{\nu}_a\}_{a\in\{0,1\}}) = [f_{1j}(\widetilde{\theta}_{1j})]^{-1}\omega_{1j}(\boldsymbol{Z}, \widetilde{\theta}_{1j}) - [f_{0j}(\widetilde{\theta}_{0j})]^{-1}\omega_{0j}(\boldsymbol{Z}, \widetilde{\theta}_{0j}).$$

Appendix B.6.1 provide details for obtaining initial estimators for the quantiles and the corresponding densities. Similar to STE, we can also obtain individual asymptotic normality for the DR estimator (3.4.7) of QTE and consistently estimate its variance.

Proposition 17 (Asymptotic normality of QTE)**.** Under the conditions in Theorem 16, when $(\vartheta^{\text{QTE}} \vee n^{-1/4})\sqrt{\log p} = o(1)$, $\delta_m = o(n^{-1/2})$ and $\sigma_j^2 := \mathbb{V}(\widetilde{\varphi}_j^{\text{QTE}}) \ge c$ for some constant $c > 0$, it holds

$$\sqrt{n}(\widehat{\tau}_j^{\text{QTE}} - \tau_j^{\text{QTE}}) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_j^2), \qquad j = 1, \ldots, p.$$

Define the sample variance $\widehat{\sigma}_j^2 = \mathbb{V}_n(\widehat{\varphi}_j^{\text{QTE}})$ for the estimated influence function $\widehat{\varphi}_j^{\text{QTE}} := \widetilde{\varphi}_j^{\text{QTE}}(\boldsymbol{Z}; \{\widehat{\boldsymbol{\theta}}_a^{\text{init}}, \widehat{\boldsymbol{f}}_a, \widehat{\pi}_a, \widehat{\boldsymbol{\nu}}_a\}_{a\in\{0,1\}})$. It further holds that $\max_{j\in[p]} |\widehat{\sigma}_j^2 - \sigma_j^2| = \mathcal{O}_{\mathbb{P}}(r_\sigma^{\text{QTE}})$ where $r_\sigma^{\text{QTE}} = (\log p)/n + \sqrt{(\log p)/n}\vartheta^{\text{QTE}} + \vartheta^{\text{QTE}}$.

Apart from the mild rate requirement on the nuisance functions, no metric entropy conditions are assumed in Theorem 16 and Proposition 17. This allows one to estimate nuisances with machine learning methods and achieve asymptotical normality with cross-fitting. While the doubly-robust estimators for QTE have also been considered by Chakrabortty et al. [23], Kallus et al. [81] for a single outcome ($p = 1$), they both require metric entropy or Donsker class conditions.

3.5 Simultaneous inference

3.5.1 Large-scale multiple testing

For a target estimand $\tau_j \in \{\tau_j^{\text{STE}}, \tau_j^{\text{QTE}}\}$, the asymptotic normality established in Corollary 13 and Proposition 17 can be utilized to test the null hypotheses $H_{0j} : \tau_j = \tau_j^*$ for $j = 1, \ldots, p$.

This implies that one can control the Type-I error of the individual tests using the statistics $t_j = \sqrt{n}(\widehat{\tau}_j - \tau_j^*)/\widehat{\sigma}_j$, with empirical variance given in Propositions 14 and 17. The confidence intervals for individual causal estimates can also be constructed. To conduct simultaneous inference, however, the tests above are too optimistic when multiple tests are of interest. Therefore, to obtain valid inferential statements, we must perform a multiplicity adjustment to control the family-wise error explicitly. This subsection provides simultaneous tests and confidence intervals for causal effects with multiple outcomes.

For $j \in [p]$, let $\varphi_{ij} = \widetilde{\varphi}_j(\boldsymbol{Z}_i)$ and $\widehat{\varphi}_{ij} = \widehat{\varphi}_j(\boldsymbol{Z}_i)$ be the influence function value and its estimate evaluated at the ith observation $\boldsymbol{Z}_i = (A_i, \boldsymbol{W}_i, \mathbf{X}_i)$, as defined in Propositions 14 and 17 for τ_j being τ_j^{STE} and τ_j^{QTE}, respectively. We require a condition from Chernozhukov et al. [30, Theorem J.1] for feasible inference.

Assumption 8 (Bounded variances and covariances)**.** There exist a constant $a, c_1 \in (0,1)$ and a set of informative hypotheses $\mathcal{A}^* \subseteq [p]$ such that $|\mathcal{A}^*| \geq ap$, $\max_{j \in \mathcal{A}^{*c}} \sigma_j^2 = o(1)$, $\min_{j \in \mathcal{A}^*} \sigma_j^2 \geq c_1$ and $\max_{j_1 \neq j_2 \in \mathcal{A}^*} |\operatorname{Corr}(\varphi_{1j1}, \varphi_{1j_2})| \leq 1 - c_1$.

When the value of σ_j is 0, the population distribution of the jth influence function is degenerated and has no variability. In Assumption 8, the first condition precludes the existence of such super-efficient estimators over $\mathcal{A}^*$, which is commonly required even in classical settings where the number of variables p is small compared to the sample size n [15]. In practice, one can use a small threshold c_n to screen out outcomes that have small variations and obtain a set of informative outcomes $\mathcal{A}_1 = \{j \in [p] \mid \widehat{\sigma}_j \geq c_n\}$.

For DR estimators derived in the previous section, the following Gaussian approximation result over a family of null hypotheses allows for data-dependent choices of the set of hypotheses and suggests a multiplier bootstrap procedure [30] for simultaneous inference.

Lemma 18 (Gaussian approximation for nested hypotheses)**.** For $\tau_j = \tau_j^{\mathrm{STE}}$ and $\vartheta = \vartheta^{\mathrm{STE}}$, suppose conditions in Proposition 14 and Assumption 8 hold. Futher assume that there exist some constants $c_2, C_2 > 0$ such that $\max\{\log(pn)^7/n, \log(pn)^2\vartheta, \sqrt{n\log(pn)}\delta_m\} \leq C_2 n^{-c_2}$. For all $\mathcal{S} \subseteq \mathcal{A}^* \subseteq [p]$, define $M_{\mathcal{S}} = \max_{j \in \mathcal{S}} |\sqrt{n}(\widehat{\tau}_j - \tau_j)/\widehat{\sigma}_j|$, $\widehat{\boldsymbol{\varphi}}_i = (\widehat{\varphi}_{ij})_{j \in \mathcal{S}}$, $\widehat{\boldsymbol{E}}_{\mathcal{S}} = n^{-1}\sum_{i=1}^n \widehat{\boldsymbol{\varphi}}_i \widehat{\boldsymbol{\varphi}}_i^\top$, and $\widehat{\boldsymbol{D}}_{\mathcal{S}} = \operatorname{diag}((\widehat{\sigma}_j)_{j \in \mathcal{S}})$. Consider null hypotheses $H_0^{\mathcal{S}}$ indexed by $\mathcal{S}$ that $\forall\, j \in \mathcal{S}$, $\tau_j = \tau_j^*$. As $m, n, p \to \infty$, it holds that

$$\sup_{H_0^{\mathcal{S}}: \mathcal{S} \subseteq \mathcal{A}^*} \sup_{x \in \mathbb{R}} |\mathbb{P}(\overline{M}_{\mathcal{S}} > x) - \mathbb{P}(\|\boldsymbol{g}_{\mathcal{S}}\|_\infty > x \mid \{\boldsymbol{Z}_i\}_{i=1}^n)| \xrightarrow{\mathrm{p}} 0,$$

where $\boldsymbol{g}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1} \widehat{\boldsymbol{E}}_{\mathcal{S}} \widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1})$. The conclusion also holds for $\tau_j = \tau_j^{\mathrm{QTE}}$ and $\vartheta = \vartheta^{\mathrm{QTE}}$ under conditions in Proposition 17 and Assumption 8.

When m is sufficiently large such that the error of derived outcomes δ_m is ignorable, the rate conditions in Lemma 18 can be satisfied if the *logarithm* of the numbers of hypotheses grows slower than $n^{\frac{1}{7}} \wedge \vartheta^{-\frac{1}{2}}$ for at least a polynomial factor of n. Lemma 18 suggests that if $\mathcal{A}_1 \subseteq \mathcal{A}^*$ only contains informative hypotheses, then the distribution of the maximal statistic $M_1 = \max_{j \in \mathcal{A}_1} |t_j|$ can be well approximated by $\boldsymbol{g}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}_{n1}^{-1} \boldsymbol{E}_{n1} \boldsymbol{D}_{n1}^{-1})$, where $\boldsymbol{E}_{n1} = n^{-1}\sum_{i=1}^n \widehat{\boldsymbol{\varphi}}_{i1} \widehat{\boldsymbol{\varphi}}_{i1}^\top$ is the sample covariance matrix with $\widehat{\boldsymbol{\varphi}}_{i1} = (\widehat{\varphi}_{ij})_{j \in \mathcal{A}_1}$ and $\boldsymbol{D}_{n1} = \operatorname{diag}((\widehat{\sigma}_j)_{j \in \mathcal{A}_1})$ is the diagonal matrix of the estimated standard deviations. This allows us to simulate the null distribution efficiently using the multiplier bootstrap procedure. To generate B bootstrap samples, for all $b = 1, \ldots, B$, we first sample n standard normal variables $\varepsilon_{11}^{(b)}, \ldots, \varepsilon_{n1}^{(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and then apply a linear transformation to obtain the multivariate normal vectors $\boldsymbol{g}_1^{(b)} = (\sqrt{n}\boldsymbol{D}_{n1})^{-1}\sum_{i=1}^n \varepsilon_{i1}^{(b)} \widehat{\boldsymbol{\varphi}}_{i1}$. It is easy to verify that $\boldsymbol{g}_1^{(1)}, \cdots, \boldsymbol{g}_1^{(B)} \overset{\text{i.i.d.}}{\sim}$

$\mathcal{N}(\mathbf{0}, \boldsymbol{D}_{n1}^{-1}\boldsymbol{E}_{n1}\boldsymbol{D}_{n1}^{-1})$ conditioned on $\{\boldsymbol{Z}_i\}_{i=1}^n$. Based on the bootstrap samples, we can estimate the upper α quantile of M_1 by $\widehat{q}_1(\alpha) = \inf\left\{x \,\middle|\, B^{-1}\sum_{b=1}^B \mathbb{1}\{\|\boldsymbol{g}_1^{(b)}\|_\infty \le x\} \ge 1-\alpha\right\}$. To test multiple hypotheses $H_{0j} : \tau_j = \tau_j^*$ for $j \in \mathcal{A}_1$, we reject those in the set $\widehat{\mathcal{A}} = \{j \in \mathcal{A}_1 \mid |t_j| > \widehat{q}_1(\alpha)\}$. The next proposition shows that the informative hypotheses can be identified, and the family-wise error rate (FWER) can be controlled.

Proposition 19 (Type-I error control)**.** For (τ_j, r_σ) being $(\tau_j^{\mathrm{STE}}, r_\sigma^{\mathrm{STE}})$ or $(\tau_j^{\mathrm{QTE}}, r_\sigma^{\mathrm{QTE}})$, suppose conditions in Lemma 18 hold . Let $\mathcal{V}^* = \{j \mid H_{0j} \text{ is false}, j = 1, \ldots, p\} \cap \mathcal{A}^*$ be the set of informative non-null hypotheses. If $\max\{r_\sigma, \max_{j\in\mathcal{A}^{*c}} \sigma_j^2\} = o(c_n)$, then as $m, n, p \to \infty$, it holds that $\lim\ \mathbb{P}(\mathcal{A}^* = \mathcal{A}_1) = 1$ and $\limsup\ \mathbb{P}(\widehat{\mathcal{A}} \cap \mathcal{V}^{*c} \neq \varnothing) \le \alpha$.

As suggested by Proposition 19, because the lower bound of informative variance in Assumption 8 is unknown, a slowly shrinking threshold c_n is needed to recover the true candidate set $\mathcal{A}^*$ and control the FWER. In practice, one can set c_n as a small value, such as 0.01, to exclude uninformative tests. If lowly expressed genes have already been excluded, thresholding may not be necessary.

3.5.2 False discovery rate control

When p is large, controlling for the false discovery proportion (FDP) or the false discovery rate (FDR) is more desirable to improve the powers when performing simultaneous testing. The FDP is the ratio of false positives to total discoveries, while the FDR is the expected value of the FDP. To control the FDP, we adopt the step-down procedure [58] to test the sequential hypotheses,

$$H_0^{(\ell)} : \forall\ j \in \mathcal{A}_\ell,\ \tau_j = \tau_j^*, \qquad \text{versus} \qquad H_a^{(\ell)} : \exists\ j \in \mathcal{A}_\ell,\ \tau_j \neq \tau_j^*, \qquad \ell = 1, 2, \ldots$$

where $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is a sequence of nested sets. The proposed multiple testing method in Algorithm 2 incorporates both the Gaussian multiplier bootstrap and step-down procedure, which aims to control the FDP exceedance rate FDX $:= \mathbb{P}(\mathrm{FDP} > c)$, the probability that FDP surpasses a given threshold c at a confidence level α. This provides a strengthened control on FDP and is asymptotically powerful, as shown in the following theorem.

Theorem 20 (Multiple testing)**.** Under the conditions of Proposition 19, consider testing multiple hypotheses $H_{0j} : \tau_j = 0$ versus $H_{aj} : \tau_j \neq 0$ for $j = 1, \ldots, p$ based on the step-down procedure with augmentation. As $m, n, p \to \infty$, the set of discoveries $\mathcal{V}$ returned by Algorithm 2 satisfies that

- (FDX) $\limsup \mathbb{P}(\mathrm{FDP} > c) \le \alpha$ where $\mathrm{FDP} = |\mathcal{V} \cap \mathcal{V}^{*c}|/|\mathcal{V}|$.
- (Power) $\mathbb{P}(\mathcal{V}^* \subset \mathcal{V}) \to 1$ if $\min_{j\in\{j\in[p]|\tau_j\neq 0\}} |\tau_j| \ge c'\sqrt{\log(p)/n}$ for some constant $c' > 0$.

Theorem 20 extend previous results by Belloni et al. [15] for many approximate means and by Qiu et al. [139] for IPW estimators to DR estimators. On the one hand, Belloni et al. [15] directly imposes assumptions on the influence functions and linearization errors, while we need to analyze the effect of nuisance functions estimation for the doubly robust estimators. On the other hand, Qiu et al. [139] requires sub-Gaussian assumptions and $\sqrt{n}$-consistency of the maximum likelihood estimation for the propensity score to establish Gaussian approximation for their proposed statistics, which does not apply to our problem setups.

Algorithm 2 Multiple testing on doubly robust estimation of treatment effects

Input: The estimated centered influence function values $\widehat{\varphi}_{ij}$, the estimated variance $\widehat{\sigma}_j^2$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The FDP exceedance threshold c and probability α, and the number of bootstrap samples B.

Initialize the iteration number $\ell = 1$, the candidate set $\mathcal{A}_1 = \{j \in [p] \mid \widehat{\sigma}_j^2 \geq c_n\}$, the set of discoveries $\mathcal{V}_1 = \varnothing$, and the statistic $t_j = \sqrt{n}(\widehat{\tau}_j - \tau_j^*)/\widehat{\sigma}_j$ for $j \in [p]$.

while not converge **do**

  Draw multiplier bootstrap samples $\boldsymbol{g}_\ell^{(b)} = (\sqrt{n}\boldsymbol{D}_{n\ell})^{-1} \sum_{i=1}^n \varepsilon_{i\ell}^{(b)} \widehat{\boldsymbol{\varphi}}_{i\ell}$, where $\varepsilon_{i\ell}^{(b)}$'s are independent samples from $\mathcal{N}(0,1)$ for $i = 1, \ldots, n$ and $b = 1, \ldots, B$.

  Compute the maximal statistic $M_\ell = \max_{j \in \mathcal{A}_\ell} |t_j|$.

  Estimate the upper α-quantile of M_ℓ under $H_0^{(\ell)} : \forall\, j \in \mathcal{A}_\ell,\ \tau_j = \tau_j^*$ by

$$\widehat{q}_\ell(\alpha) = \inf \left\{ x \,\middle|\, \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\|\boldsymbol{g}_\ell^{(b)}\|_\infty \leq x\} \geq 1 - \alpha \right\}.$$

  if $M_\ell > \widehat{q}_\ell(\alpha)$ **then**

    Set $j_\ell = \operatorname{argmax}_{j \in \mathcal{A}_\ell} |t_j|$, $\mathcal{A}_{\ell+1} = \mathcal{A}_\ell \setminus \{j_\ell\}$, and $\mathcal{V}_{\ell+1} = \mathcal{V}_\ell \cup \{j_\ell\}$.

  else

    Declare the treatment effects in $\mathcal{A}_\ell$ are not significant and stop the step-down process.

  end if

  $\ell \leftarrow \ell + 1$.

end while

Augmentation: Set $\mathcal{V}$ to be the union of $\mathcal{V}_\ell$ and the $\lfloor |\mathcal{V}_\ell| \cdot c/(1-c) \rfloor$ elements from $\mathcal{A}_\ell$ with largest magnitudes of $|t_j|$.

Output: The set of discoveries $\mathcal{V}$.

3.6 Simulation

We consider a simulation setting with $p = 8000$ genes and generate an active set of genes $\mathcal{V}^* = \mathcal{A}^* \subset [p]$ with size 200. We draw covariates $\boldsymbol{W} \in \mathbb{R}^d$ with i.i.d. $\mathcal{N}(0,1)$ entries and the treatment A follows a logistic regression model with probability $\mathbb{P}(A = 1 \mid \boldsymbol{W}) = 1/(1 + \exp(\mathbf{1}_d^\top \boldsymbol{W}/(d + 1)))$. Then, we generate the counterfactual gene expressions. For a gene j, the single-cell gene expression $X_j(0)$ is drawn from a Poisson distribution with mean $\lambda_j = \exp(\boldsymbol{W}^\top \boldsymbol{b}_j) \in \mathbb{R}$ where the entries of both the coefficients $\boldsymbol{b}_j \in \mathbb{R}^d$ with 1 as the first entry and the remaining entries independently drawn from $\mathcal{N}(0, 1/4)$. The gene expressions $X_j(1)$ for $j \notin \mathcal{V}^*$ are generated from the same model, while for gene $j \in \mathcal{V}^*$, we consider two treatment mechanisms that favor the mean-based and quantile-based tests, respectively; see Appendix B.6.2 for more details about the data generating processes.

Next, we draw m observations $\boldsymbol{X}_1(A), \ldots, \boldsymbol{X}_m(A)$ independently, which are summed up as the overall gene expression $\widetilde{\boldsymbol{Y}}(A)$. Then, the observed gene expression matrix is given by $\mathbf{X} = A\mathbf{X}(1) + (1-A)\mathbf{X}(0)$ and analogously $\widetilde{\boldsymbol{Y}} = A\widetilde{\boldsymbol{Y}}(1) + (1-A)\widetilde{\boldsymbol{Y}}(0)$. We then draw n independent observed samples $\{(A_i, \boldsymbol{W}_i, \mathbf{X}_i, \widetilde{\boldsymbol{Y}}_i)\}_{i=1}^n$. The parameters are set to be $d = 5$, $m = 100$, $n \in \{100, 200, 300, 400\}$. For nuisance function estimation, we employ Logistic regression to estimate the propensity score and Poisson generalized linear model (GLM) with the log link to estimate the mean regression functions. For quantile-based methods, the initial estimators of the quantile

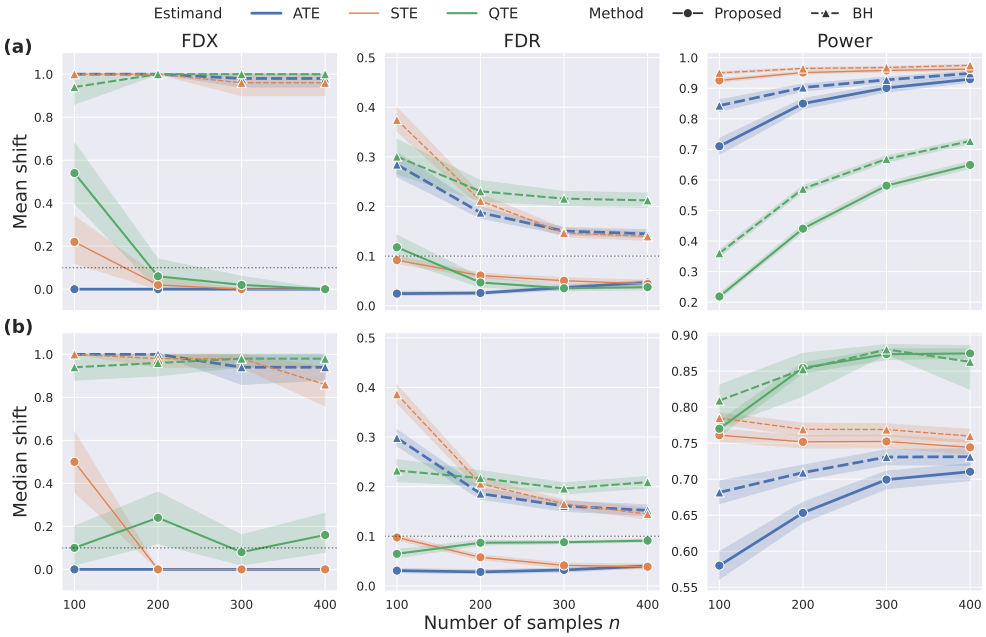

Figure 3.2: Simulation results of the hypothesis testing of $p = 8000$ outcomes based on different causal estimands and FDP control methods for detecting differential signals under **(a)** mean shifts and **(b)** median shifts averaged over 50 randomly simulated datasets without sample splitting. The gray dotted lines denote the nominal level of 0.1.

and density are described in Appendix B.6.1.

To quantify the performance of different test statistics and multiple testing procedures, we compare the empirical FDX, FDR, and power. We aimed to control FDX over 0.1 at 0.05, namely $\mathbb{P}(\text{FDP} > 0.1) \leq 0.05$. We also compared with the Benjamini-Hochberg (BH) procedure with targeting FDR controlled at 0.05. The experiment results are summarized in Figure 3.2 without sample splitting and in Figure B.62 of Appendix B.6.2 with cross-fitting. As shown in Figure 3.2, in the high signal-noise ratio (SNR) setting, the proposed method controls both FDX and FDR at the desired level for all three causal estimands when the sample size is relatively large, i.e., $n > 100$. On the other hand, the BH procedure fails to control the FDX and FDP at any sample sizes because the p-values are not close to uniform distribution (see Figure B.61), though the gaps of FDP become smaller when the sample size gets larger. Though the BH procedure has lower FDP with sample splitting (see Figure B.62), it still fails to control FDX for all estimands. This indicates that the proposed multiple testing procedure consistently outperforms the BH procedure by correctly accounting for the dependencies among the test statistics and providing valid statistical error control.

In the low SNR setting, we see that the quantile-based estimand has larger powers than mean-based tests, which is expected because of the designed data-generating process. Such a low SNR scenario is often encountered with scRNA-seq data. In this case, the proposed method still has better control of both FDX and FDR compared to the BH procedure. Although the QTE test is slightly anti-conservative regarding FDX, it still controls the FDR well. Furthermore, standardized tests are more powerful than unstandardized estimands, especially when the sample size is small. Overall, the results in Figure 3.2 demonstrate the valid FDP control of the proposed multiple testing procedure for various causal estimands and suggest that testing based on different

estimands could be helpful in different scenarios. In contrast, the commonly used BH procedure in genomics may be substantially biased due to the complex dependency among tests.

3.7 causarray

3.7.1 Doubly-robust counterfactual imputation and inference

Our objective is to determine whether a gene is causally affected by a "treatment" variable after controlling for other technical and biological covariates, which may affect the treatment and outcome variables. Here, we use the term treatment generally; in the narrow sense, it can mean genetic and/or chemical perturbations [79, 118], such as CRISPR-CAS9, and, more broadly, it can mean the phenotype of a disease [134]. We acknowledge that while many differentially expressed genes can be considered a result of disease status, for most late-onset disorders, a smaller fraction of genes could have initiated disease phenotypes. Our method aims to determine the direct effects of treatments on modulated gene expression outcomes.

In observational data, the response variable can be confounded by measured and unmeasured biological and technical covariates, making it difficult to separate the treatment effect from other unknown covariates. As a consequence, it is challenging to draw causal inferences; even tests of association may lead to an excess of false discoveries and/or low power. Fortunately, the potential outcomes framework [143, 152] formulates general causal problems in a way that allows for the treatment effect to be separated from the effects of other variables. However, even this framework is challenged by unmeasured covariates. Before introducing our method for estimating unmeasured confounders, we first outline the general potential outcomes framework.

Consider a study in which Y is the response variable and A is the binary treatment variable for an observation. In the potential outcomes framework, $Y(a)$ is the outcome that we would have observed if we set the treatment to $A = a$. Naturally, we can only observe one of the two potential outcomes for each observation, so

$$Y = \mathbb{1}\{A = 1\}Y(1) + \mathbb{1}\{A = 0\}Y(0),$$

In the context of a case-control study of a disease, this would answer the question: What is the expected difference in gene expression if an individual had the disease (case, $A = 1$) versus if they did not (control, $A = 0$)?

Doubly robust methods provide a powerful tool for estimating potential outcomes in observational studies where randomization is not possible [143, 152]. Specifically, we estimate two key quantities: (1) $\mu_a(X)$, the mean response of the outcome variable conditional on treatment $A = a$ and covariates $X = x$, and (2) $\pi_a(X)$, the propensity score, which is defined as the probability of receiving treatment $A = a$ given covariates X, i.e., $\pi_a(X) = \mathbb{P}(A = a \mid X)$. Using these estimates, we compute potential outcomes as

$$\widehat{Y}(a) = \frac{\mathbb{1}\{A = a\}}{\widehat{\pi}_a(X)}(Y - \widehat{\mu}_a(X)) + \widehat{\mu}_a(X).$$

The doubly robust estimator's name comes from the fact that it provides a consistent estimate as long as *either* the outcome model, $\mu_a(X)$, or the propensity score model, $\pi_a(X)$, is correctly specified. Given this estimate, we can easily perform downstream inference tasks such as computing log fold change (LFC), and testing for causal effects on gene expressions (fig. 3.3a). An advantage of this approach is that counterfactual imputation denoises/balances gene expression under

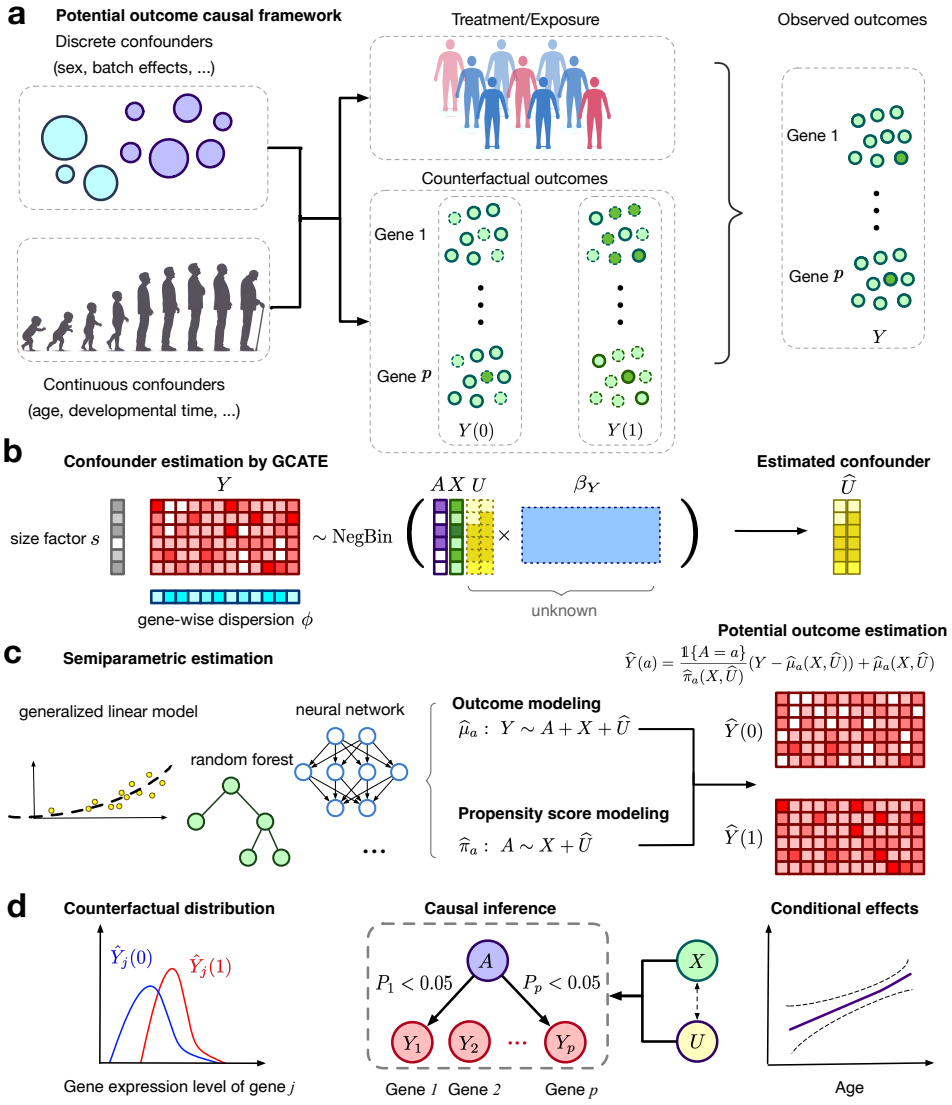

Figure 3.3: Overview of the proposed causarray method. a, Illustration of the data generation process for pseudo-bulk and single-cell data. **b,** The gene expression matrix, Y, is linked to the treatment, A, measured covariates, X, and confounding variables, U, via a GLM model. The cell-wise size factor, s, and gene-wise dispersion parameter, ϕ, are estimated from the data, and the unmeasured confounder U is estimated by $\widehat{U}$ through the augmented GCATE method. **c,** Generalized linear models and flexible machine learning methods including random forest and neural network can be applied for outcome modeling ($\mathbb{E}[Y \mid A = a, X, \widehat{U}] = \widehat{\mu}_a(X, \widehat{U})$) and propensity modeling ($\mathbb{P}(A = a \mid X, U) = \widehat{\pi}_a(X, \widehat{U})$) The estimated outcome and propensity score functions give rise to the estimated potential outcomes for each cell and each gene. **d,** Downstream analysis includes contrasting the estimated counterfactual distributions, performing causal inference, and estimating the conditional average treatment effects.

two different conditions. Additionally, having access to estimated potential outcomes facilitates downstream analyses such as estimating causal effects conditional on measured confounders like age.

A key step in these types of analyses is estimating unmeasured confounders. To adjust for confounding, factor models were popularized in surrogate variable analysis literature and have since been widely adopted in bulk gene expression studies [96]. Recently, we extended this approach to single-cell RNA-seq data using generalized linear models that better accommodate pseudobulk and single-cell outcome variables [48]. Using this generalized factor analysis approach, we estimate unmeasured confounders U alongside potential outcomes (fig. 3.3b-c), enabling direct estimation of downstream quantities such as LFC (fig. 3.3d).

3.7.2 causarray applied to an in vivo Perturb-seq study reveals causal effects of ASD/ND genes

An integrative analysis of multiple single perturbations Autism spectrum disorders and neurodevelopmental delay (ASD/ND) represent a complex group of conditions that have been extensively studied using genetic approaches. To investigate the underlying mechanisms of these disorders, researchers have employed scalable genetic screening with CRISPR-Cas9 technology [79]. Frameshift mutations were introduced in the developing mouse neocortex in utero, followed by single-cell transcriptomic analysis of perturbed cells from the early postnatal brain [79]. These in vivo single-cell Perturb-seq data allow for the investigation of causal effects of a panel of ASD/ND risk genes. We analyze the transcriptome of cortical projection neurons (excitatory neurons) perturbed by one risk gene or a non-targeting control perturbation, which serves as a negative control.

Unmeasured confounders, such as batch effects and unwanted variation, are likely present in this dataset due to the batch design being highly correlated with perturbation conditions (fig. B.64ab). Additionally, the heterogeneity of single cells assessed in vivo introduces further complexity. These confounding factors may reduce statistical power for gene-level differential expression (DE) tests, as noted in the original study [79], which instead focused on gene module-level effects. To address this limitation, we apply causarray to incorporate unmeasured confounder adjustment and conduct a more granular analysis at the single-gene level. This approach enables us to uncover nuanced genetic interactions and causal effects that may provide deeper insights into the etiology of ASD/ND.

Functional analysis Gene module-level analyses have been shown to provide greater statistical power for detecting biologically meaningful perturbation effects when fewer cells are available [79]. The original study adopted this approach but relied on a linear model rather than a negative binomial model, potentially limiting its ability to detect broader signals at the individual gene level. Here, we compare causarray with RUV and DESeq2 (without confounder adjustment) to identify significant genes and enriched gene ontology (GO) terms associated with various perturbations. The number of latent factors is set as 10, according to the joint-likelihood-based information criterion (fig. B.65a).

In terms of significant gene detection, causarray identifies a comparable number of significant genes to RUV across most perturbations, while DESeq2 consistently detects fewer significant genes (fig. 3.4a). The variation in significant detections across different perturbed genes suggests distinct biological impacts of each knockout. Functional analysis focuses on enriched GO terms on the DE genes under each perturbation condition where discrepancies arise between causarray and other methods. Genes identified by causarray are enriched for biologically relevant GO terms with clear clustering patterns (fig. 3.4b-c, fig. B.64c). In contrast, RUV shows less distinct clustering and enrichment patterns.

Notably, while RUV identifies GO terms related to ribosome processes previously implicated in ASD studies [107], these findings remain controversial. Some argue that dysregulation in translation processes and ribosomal proteins may reflect secondary changes triggered by expression alterations in synaptic genes rather than direct causal effects [60]. In contrast, GO terms identified by causarray align more closely with the expected causal effects of ASD/ND gene perturbations [54, 93].

To further validate these findings, we examine the perturbation condition for *Satb2*, which yields the largest number of significant genes identified by both methods (adjusted P value < 0.1) and exhibits significant different estimated propensity scores (fig. B.65b). *Satb2* is known to play critical roles in neuronal development, synaptic function, and cognitive processes [173, 184]. Using causarray, we detect enrichment for GO terms directly related to neuronal function and development, such as "regulation of neuron projection development," "regulation of synapse structure or activity," and "synapse organization" (fig. 3.4d). These findings are consistent with *Satb2*'s established roles in neuronal development and synaptic plasticity [61, 73]. On the other hand, RUV identifies enrichment for terms related to mitochondrial function and energy metabolism, such as "mitochondrial electron transport," "cellular respiration," and "ATP synthesis" (fig. 3.4e). While these processes are important for general cellular function, they are less directly relevant to *Satb2*'s primary biological roles.

Overall, this analysis demonstrates that causarray provides greater specificity in detecting biologically meaningful causal effects of gene perturbations. Its ability to disentangle confounding influences while preserving relevant biological signals highlights its effectiveness in analyzing complex genomic datasets.

3.7.3 causarray reveals causally affected genes of Alzheimer's disease in a case-control study

An integrative analysis of excitatory neurons We analyze three Alzheimer's disease (AD) single-nucleus RNA sequencing (snRNA-seq) datasets: a transcriptomic atlas from the Religious Orders Study and Memory and Aging Project (ROSMAP) [116] and two datasets from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium [55], which include samples from the middle temporal gyrus (MTG) and prefrontal cortex (PFC). Our objective is to compare the performance of causarray and RUV in pseudo-bulk DE tests of AD in excitatory neurons.

To evaluate the validity, we perform a permutation experiment on the ROSMAP-AD dataset by permuting phenotypic labels. Ideally, no significant discoveries should be made under this null scenario. However, RUV produces a large number of false discoveries, with its performance deteriorating as the number of latent factors increases. In contrast, causarray effectively controls the false discovery rate (FDR), producing minimal false positives (fig. 3.5a). Additionally, we assess coherence across datasets by examining effect sizes in SEA-AD (MTG) and SEA-AD (PFC). Effect sizes estimated by causarray exhibit higher consistency across varying q-value cutoffs compared to RUV (fig. 3.5b, fig. B.66b). When inspecting DE genes across all three AD datasets, causarray identifies more consistent discoveries than RUV (fig. 3.5c), highlighting its robustness in detecting causally affected genes.

Functional analysis We further compare functional enrichment results between causarray and RUV using gene ontology (GO) terms associated with DE genes. Across the three datasets, causarray identifies 165 common GO terms, significantly more than the 60 identified by RUV

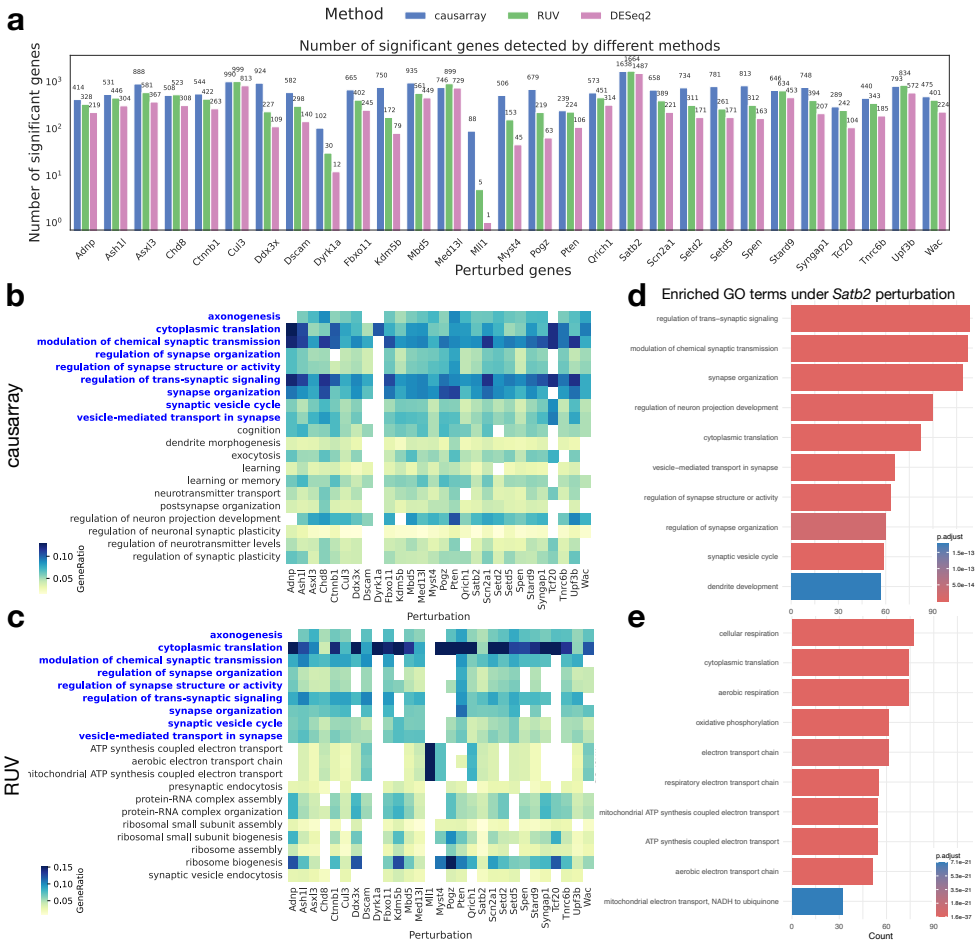

Figure 3.4: Statistical test results of the effects of CRISPR perturbation on gene expression in excitatory neuron data. a, Number of significant genes detected under all perturbations using three different methods. The detection threshold for significant genes is FDR< 0.1 for all methods. **b-c,** Heatmaps of GO terms enriched (adjusted P value < 0.05, $q < 0.2$) in discoveries from causarray and RUV, respectively, where the common GO terms are highlighted in blue. Only the top 20 GO terms that have the most occurrences in all perturbations are displayed. **d-e,** Barplots of GO terms enriched in discoveries under *Satb2* perturbation from causarray and RUV, respectively.

(fig. 3.5d). Both methods detect GO terms relevant to neuronal development and synaptic functions, which are critical for understanding AD pathology. However, causarray shows distinct enrichment in categories such as "positive regulation of cell development" and "negative regulation of cell cycle', reflecting its increased sensitivity to synaptic and neurotransmission-related processes. In contrast, RUV's results exhibit more dataset-specific enrichments, such as biosynthetic processes in SEA-AD (PFC), apoptotic processes in SEA-AD (MTG), and catabolic processes in ROSMAP-AD (fig. B.66c). These findings suggest that causarray captures more generalizable biological signals across datasets.

Both methods identify overlapping top functional categories related to key biological processes associated with AD pathology (fig. B.66e). However, causarray associates a larger number of genes with these categories, identifying 3393 DE genes compared to 3187 for RUV (fig. 3.5c).

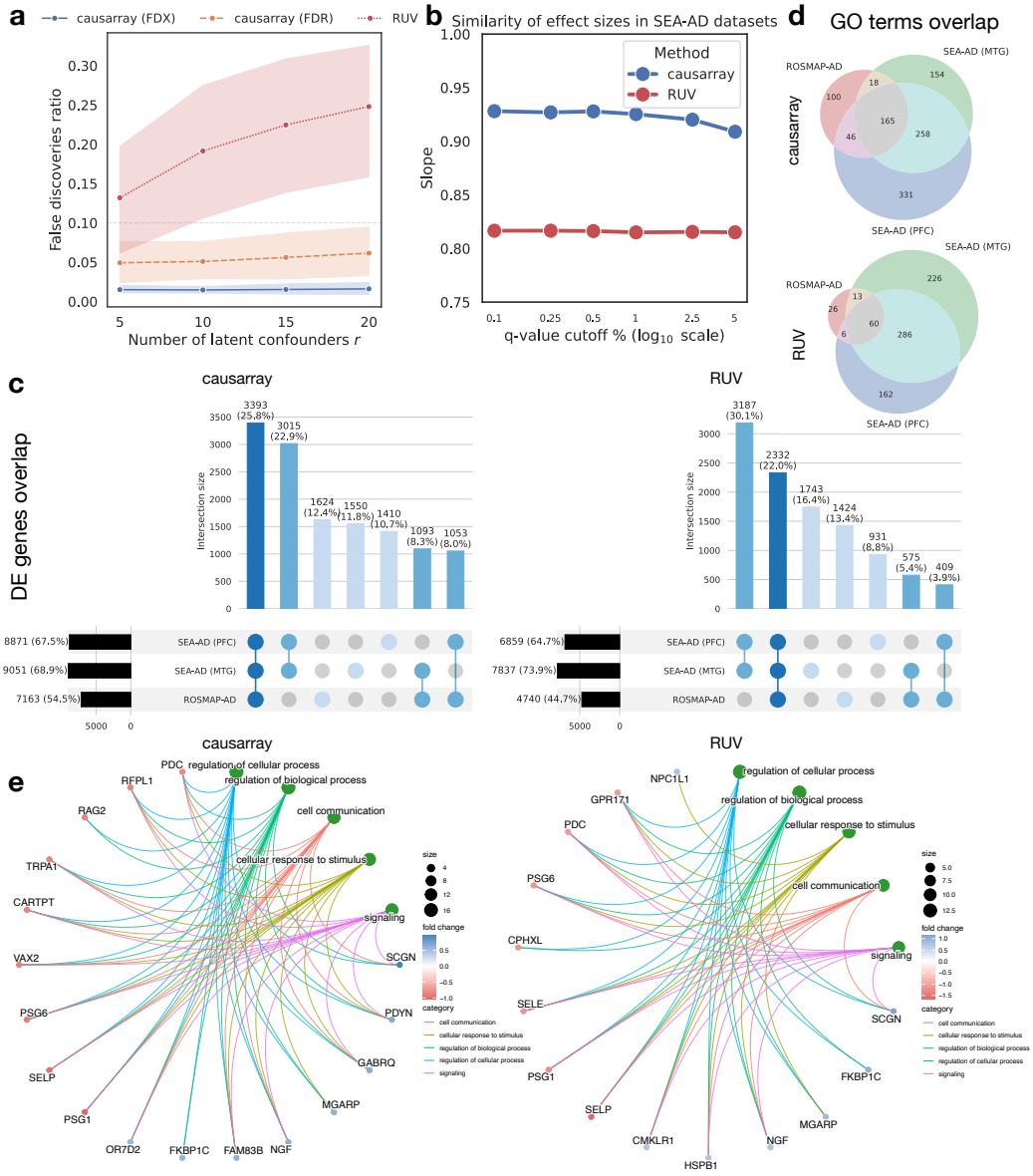

Figure 3.5: Comparison of DE genes discovered by causarray and RUV on excitatory neurons for Alzheimer's disease. a, The ratio of false discoveries to all 15586 genes of DE test results with permuted disease labels on the ROSMAP-AD dataset. Three methods, causarray with FDX control, causarray with FDR control, and RUV with FDR control, are compared. Data are presented as mean values ± s.d. **b,** The similarity of estimated effect sizes on SEA-AD MTG and PFC datasets. The slope is estimated from linear regression of effect sizes on the PFC dataset against those on the MTG dataset. **c,** DE genes by causarray and RUV over 15586 genes (adjusted P value < 0.1). **d,** Venn diagram of associated GO terms from causarray and RUV (adjusted P value < 0.05, $q < 0.2$). **e,** Considering only the top 50 positively regulated and the top 50 negatively regulated DE genes from causarray and RUV, we map them to the top 5 biological processes (the green nodes).

Additionally, causarray reveals 165 common GO terms across the three datasets, significantly more than the 60 identified by RUV (fig. 3.5d). The visualization of the discovered networks, as

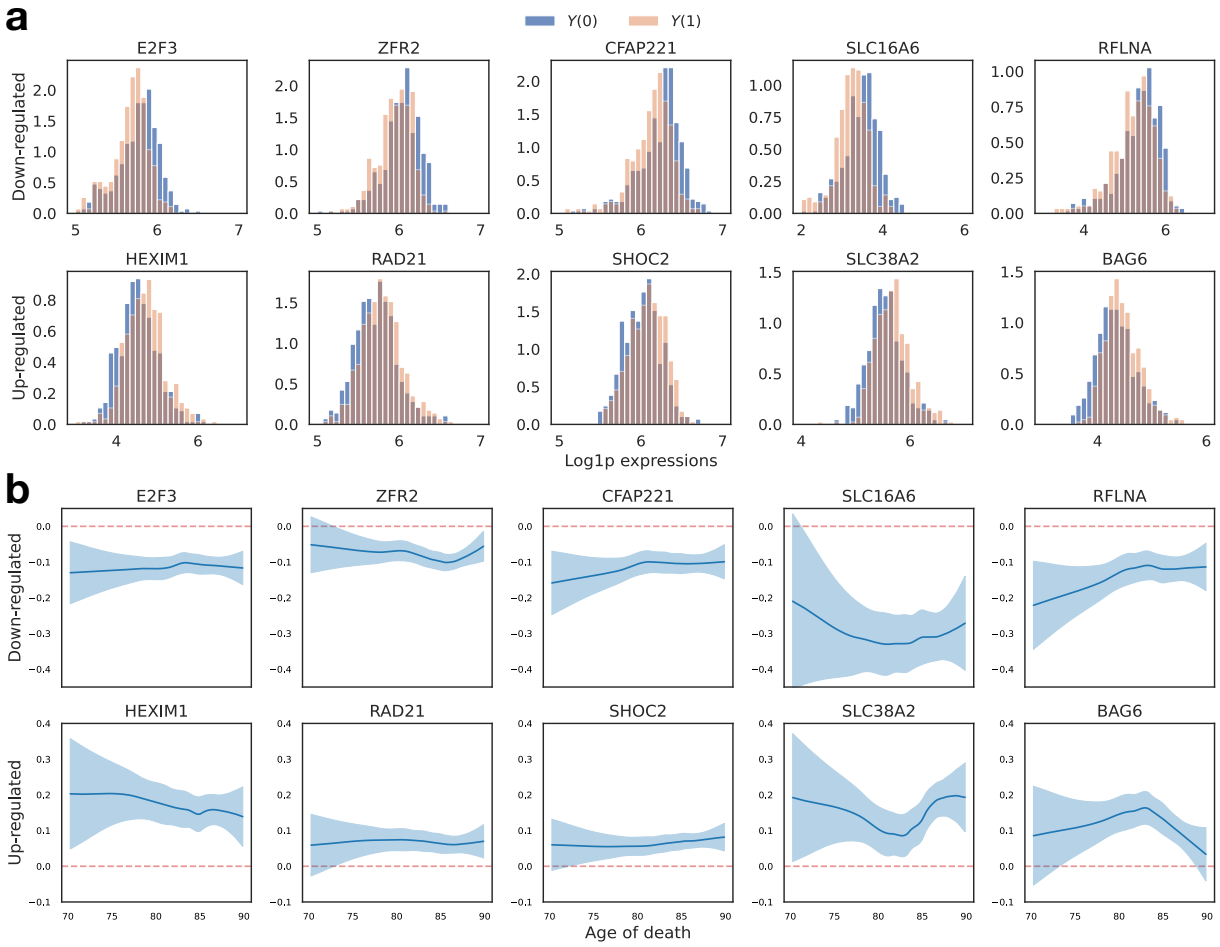

Figure 3.6: Results of DE analysis of 10 selected genes by causarray. The top 5 up-regulated and top 5 down-regulated genes in estimated LFCs (adjusted P value < 0.05) are visualized. **a,** Estimated counterfactual distributions. The values are shown in the log scale after adding one pseudo-count. **b,** Estimated log-fold change of treatment effects, conditional on age for selected genes. The center lines represent the mean of the locally estimated scatter plot smoothing (LOESS) regression, and the shaded area represents a 95% confidence interval at each value of age.

defined as the top 5 GO terms and associated genes included in the top 100 DE gene discoveries, further highlights the enhanced sensitivity and comprehensiveness of causarray. Specifically, the causarray network contains 17 gene nodes and 81 edges, compared to 14 gene nodes and 57 edges in the RUV network (fig. 3.5e). This greater interconnectedness in the larger causarray network suggests a more intricate and informative representation of underlying biological relationships, emphasizing its ability to capture broader and more relevant genetic factors associated with AD pathology.

Counterfactual analysis The counterfactual framework employed by causarray enables downstream analyses that directly utilize estimated potential outcomes. By examining counterfactual distributions for significant genes (fig. 3.6a), we observe distinct shifts in expression levels between treatment ($Y(1)$) and control ($Y(0)$) groups. Downregulated genes show a shift toward

lower expression levels under disease conditions, while upregulated genes exhibit increased expression. Conditional average treatment effects (CATEs) reveal age-dependent trends for these genes (fig. 3.6b). For example, upregulated genes such as *SLC16A6* and *RFLNA* show stronger effects at extreme ends of the age distribution, while others like *SLC38A2* and *BAG6* display nuanced changes across the aging spectrum.

These findings align with prior studies highlighting the roles of specific genes in aging-related processes. For instance, *ZFR2*, *RFLNA*, *BAG6*, and *RAD21* have been implicated in chromatin remodeling, synaptic plasticity, and cellular stress responses critical for aging and neurodegeneration [67, 83, 94, 129]. While nonparametric fitted curves exhibit wider uncertainty bands, particularly at the boundaries, which can be observed here, the significant trends observed for key genes highlight their potential relevance in AD pathology. Overall, these results demonstrate that causarray provides nuanced insights into age-dependent gene regulation mechanisms while maintaining robust control over confounding influences.

Chapter 4

Assumption-Lean Post-Integrated Inference with Negative Control Outcomes

Material in this chapter first appeared as Du et al. [45].

4.1 Introduction

In the big data era, integrating information from multiple heterogeneous sources has become increasingly crucial for achieving larger sample sizes and more diverse study populations. The applications of data integration are in a variety of fields, including but not limited to, causal inference on heterogeneous populations [157], survey sampling [179], health policy [133], retrospective psychometrics [69], and multi-omics biological science [39]. Data integration methods have been proposed to mitigate the unwanted effects of heterogeneous datasets and unmeasured covariates, recovering the common variation across datasets. However, a critical and often overlooked question is whether reliable statistical inference can be made from integrated data. Directly performing statistical inference on integrated outcomes and covariates of interests fails to account for the complex correlation structures introduced by the data integration process, often leading to improper analyses that incorrectly assume the corrected data points are independent [101].

While data integration is broadly utilized in various fields, our paper focuses on a challenging scenario with the presence of high-dimensional outcomes. Particularly in the context of genomics, experimental constraints often necessitate the collection of data in multiple batches [111, 112]. Batch correction and data integration methods are commonly used in genomics to recover the *low-dimensional embeddings* or manifolds of each observation from the *high-dimensional outcomes*. The naive approach uses a batch indicator as a covariate in a regression model for inference, which may not be sufficient for adjusting for batch effects and unmeasured covariates [101]. Instead, two-step methods are commonly employed in practice as a separate data preprocessing step to produce integrated data, which can then be utilized for downstream inference. For instance, design-based methods, such as Combat [80] and BUS [112], combine the batch or unknown subtype indicator into hierarchical Bayesian models and provide location and scale correction. Additionally, design-free methods, including RUV [56] and SVA [99] directly estimate the latent confounding factors, and users can use the estimated latent variables as extra covariates for the downstream inference. These methods apply to samples that share the same underlying biological variability, which is

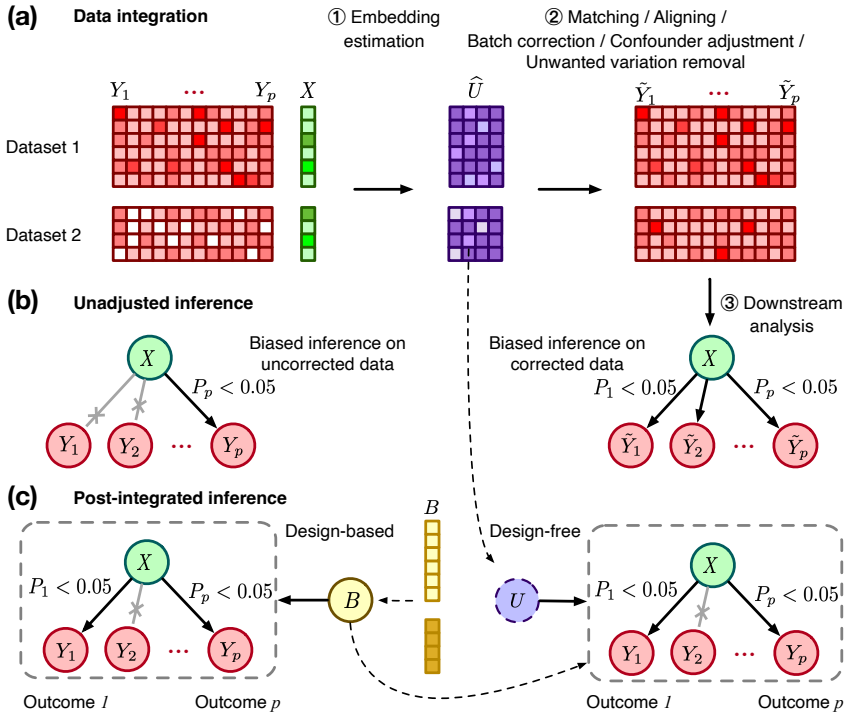

Figure 4.1: Overview of the post-integrated inference problem. **(a)** Data integration utilizes multiple outcomes $Y = (Y_1, \ldots, Y_p)^\top$ and covariate X of interest to estimate the embeddings $\widehat{U}$, and provides integrated outcomes $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_p)^\top$ for downstream analysis. **(b)** Inference on the direct associations between Y_j's and X, and those between $\widetilde{Y}_j$'s and X may be biased because of batch effects and observational dependency induced by data integration processes, respectively. **(c)** Post-integrated inference includes two strategies: the design-based approach that includes a batch indicator through a statistical model and the design-free approach that first estimates the latent embeddings and then treats them as extra covariates for downstream inference (the batch indicator can also be used as an observed confounder), where the latter is our focus.

our focus in this paper; see Figure 4.1 for an illustration.

Despite different procedures and output formats, nearly all batch correction methods utilize information from multiple outcomes to estimate and align the underlying "embeddings" of observations. This approach is closely related to unmeasured confounder adjustment, particularly when each observation is viewed as a single dataset. Over the past decades, researchers have explored various methods to address unmeasured confounders in statistical analysis. In the presence of multiple outcomes, deconfounding techniques primarily employ two strategies: incorporating known negative control outcomes or leveraging sparsity assumptions [175, 189]. Additionally, a line of research on proximal causal inference uses both negative control outcomes and/or exposures for deconfounding [124]; see a review of related work in Appendix C.1. This paper focuses specifically on the negative control approach in the context of multiple outcomes, where the goal is to directly estimate and adjust for latent factors that may confound the treatment outcome relationships.

Mathematically, a high-dimensional outcome vector $Y \in \mathbb{R}^p$ is often related to a covariate vector $X \in \mathbb{R}^d$ and an unobserved low-dimensional latent vector $U \in \mathbb{R}^r$. Here, X includes variables such as disease status or treatment, and U, frequently referred to as the embedding

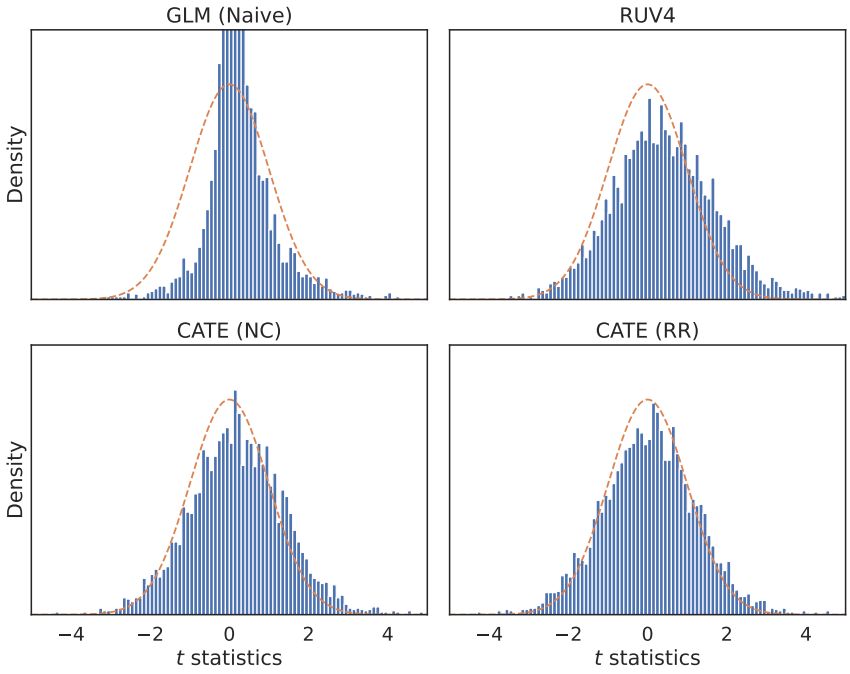

Figure 4.2: Histogram of t-statistics of *PTEN* perturbation on 8320 cells and 4163 genes by four different confounder adjustment methods. The orange dashed curves represent the density of standard normal distribution. See Section 4.5 for more details about the methods and experiment setting.

vector, captures both the batch effects and the unmeasured covariates. Both of them serve as a compact representation of the outcome Y, with the dimensionality of the outcome space being significantly larger than that of the covariate and latent space, i.e., $p \gg d$ and $p \gg r$. Differences in how data are collected across datasets can result in shifts or distortions in the distribution of unobserved variable U, and can potentially affect the distribution of X as well. Our primary interest lies in the direct associations or causal relationships between the outcome Y_j and the covariate X for $j = 1, \ldots, p$, after adjusting for the difference induced by unwanted variation U. When X and U are independent, the problem would be trivial because the direct effects can be estimated by regressing Y_j's on X. However, when X and U are dependent, the direct regression approach targets the total effects and provides a biased estimate of the direct effects. Hence, proper data integration methods need to estimate U for integrating the outcomes from different sources and for multiple hypotheses testing.

Although two-step procedures are widely favored by practitioners, it is evident that the risk of making mistakes is propagated by the two steps. Specifically, the estimation of latent embeddings U and the subsequent statistical inference are both contingent on the assumptions made by their respective models. If either model is misspecified, the final inference results can be significantly biased. For instance, varying choices of hyperparameters, such as the latent dimension, can affect the accuracy of the first-stage estimation. It is, therefore, critical to understand whether such approaches work in more general settings and how to remedy these existing post-integrated inference methods under possible misspecification.

In this paper, we rigorously investigate the validity of statistical inference on integrated data, focusing particularly on the use of negative control outcomes to ensure reliable inference. Our aim is to analyze the validity of two-step post-integrated inference under minimal assumptions

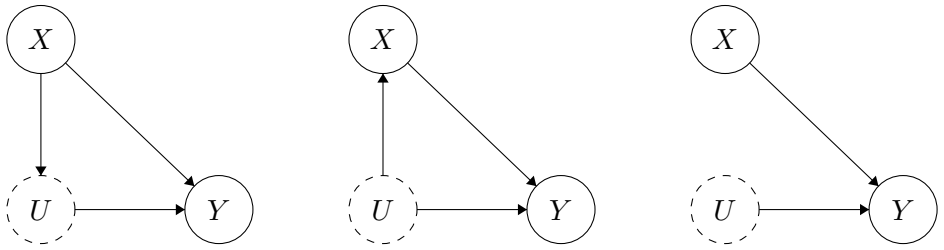

Figure 4.3: Batch correction where the latent embedding U is (a) U is a mediator that contributes to the indirect effect from X to Y; (b) a confounder that affects both X and Y; and (c) a moderator that interacts with an independent variable X of interest to influence an outcome Y, but is not on the causal pathway.

about the data-generating processes. Further, we aim to provide a framework that not only ensures effective batch correction but also maintains the integrity and reliability of statistical inference, addressing two key challenges using flexible machine learning algorithms. This will allow researchers to retain the statistical power of their analyses while providing greater confidence in the validity of their inferences from integrated data.

Post-Integrated inference To demonstrate the challenges in post-integrated inference, we analyze high-throughput single-cell CRISPR data from a study on autism spectrum disorder-related gene perturbations and their effects on neuronal differentiation (Section 4.5). In this example, one cell can be viewed as a single dataset, where the heterogeneity among cells may not be fully explained by observed covariates. Our analysis focuses on testing nonlinear associations between 4163 genes and *PTEN* perturbation after accounting for covariates in neural development and unwanted variations from heterogeneous observations.

Figure 4.2 illustrates t-statistic distributions from four different methods. The unadjusted inference method yields overly conservative test statistic distributions compared to the expected $\mathcal{N}(0,1)$ distribution. While batch correction and confounder adjustment methods produce distributions closer to the standard normal, some show anti-conservative tendencies. Importantly, only about half of the significant tests (p-value<0.05) are consistent across the three confounder adjustment methods, raising concerns about their reliability. This inconsistency stems from varied model assumptions and algorithms tailored to specific data models, which may be misspecified for sparse single-cell data or due to inaccurate estimation of the number of latent factors.

The goal of this paper is to construct a robust statistical framework that uses embeddings from existing data integration methods to mitigate misspecification issues, ensuring valid statistical inference and enhancing current post-integrated inference methodologies.

Main contributions Our work makes several key contributions. First, in Section 4.2, we derive nonparametric identification conditions using negative control outcomes (Section 4.2.1), enhancing causal interpretations and forming the basis for our post-integrated inference (PII) method. In Section 4.2.2, introduce a robust and assumption-lean framework for post-integrated inference that effectively addresses hidden mediators, confounders, and moderators (Figure 4.3). This framework ensures reliable statistical inference despite possible confounding from batch effects and data heterogeneity. It eliminates confounding ambiguity (Remark 8), leverages negative control outcomes for accurate embedding estimation (Remark 9), and exhibits resilience to model misspecification, supporting model-free inference (Remark 10).

Our second contribution in Section 4.3 analyzes the statistical error in target estimands using estimated embeddings. In Section 4.3.1, we use martingale interpretations to assess the bias caused by these embeddings. Under regularity conditions, we show in Theorem 22 that the bias of the projected target estimand with estimated embeddings is primarily determined by the L_2-norm of the embedding estimation error, up to an invertible transformation. Furthermore, Lemma 23 shows that bias in linear models can be deterministically evaluated using the operator norm of projection matrices, regardless of latent dimensionality.

Our third main contribution, discussed in Section 4.3.2 and appendix C.5.3, involves developing efficient semiparametric inference methods for estimands with estimated covariates under both linear and nonlinear functions. These methods advance the assumption-lean approach of Vansteelandt and Dukes [171] to accommodate multiple treatments and outcomes. Specifically, we derive finite-sample linear expansions for direct effect estimands (Theorem 24), provide a uniform concentration bound for residuals, and establish the asymptotic distribution for both linear and nonlinear effects under mild assumptions (Corollary 25, Theorem C.5.1) with triangular arrays. These results are essential for establishing guarantees related to multiple testing with high-dimensional outcomes (Proposition 26).

4.2 Post-Integrated inference

4.2.1 Nonparametric identification with negative control outcomes

In this section, we consider post-integrated inference with negative control outcomes. Similar to the causal inference analysis with observational data [71, 84], we consider the case when the latent variable U is a confounder as in Figure 4.3(b). Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{U} \subseteq \mathbb{R}^r$ be the support of X and U, respectively. We use f to denote a generic (conditional) probability density or mass function and require causal assumptions on the observational data (X, U, Y) and counterfactual outcome $Y(x)$ when X is interpreted as treatment.

Assumption 9.
1. Consistency: when $X = x$, $Y = Y(x)$.
2. Positivity: $f(x \mid u) > 0$ for all $u \in \mathcal{U}$.
3. Latent ignorability: $X \perp\!\!\!\perp Y(x) \mid U$ for all $x \in \mathcal{X}$.

Assumption 91 requires that no interference among the subjects, meaning that a subject's outcome is affected by its treatment but not by how others are treated. Assumption 92 suggests that $X = x$ can be observed at any confounding levels of U with a positive probability. Assumption 93 ensures that the treatment assignment is fully determined by the confounder U. These assumptions are required to estimate the counterfactual distribution of $Y(x)$ with observed variables (X, U, Y) by the g-formula $f_{Y(x)}(y) = \int f(y \mid u, x) f(u) \, \mathrm{d}u$. In our problem, because U is not observed, all information contained in the observed data is captured by $f(y, x)$, and one has to solve for $f(y, x, u)$ or equivalently $f(u \mid y, x)$ from the integral equation:

$$f(y, x) = \int f(y, x, u) \, \mathrm{d}u. \tag{4.2.1}$$

In general, the joint distribution $f(y, x, u)$ cannot be uniquely determined. With an auxiliary variable Z, the approach by Miao et al. [125, Theorem 1] identifies the treatment effect from any admissible[1] distribution under exclusion restriction, equivalence, and completeness assumptions.

[1] A joint distribution $\widetilde{f}(y, x, u)$ is admissible if it conforms to the observed data distribution $f(y, x)$, that is, $f(y, x) = \int \widetilde{f}(y, x, u) \, \mathrm{d}u$.

Because the negative control outcomes can also be viewed as a non-differentiable proxy of the confounder, their result also applies to our problem if taking $Z = Y_{\mathcal{C}}$; however, when restricting to negative control outcomes, we can extend their approach on the identification of the counterfactual distributions with weaker assumptions. Now, we modify it as follows.

To present our first result on identification with negative control outcomes, we let $f(y, x, u; \alpha)$ denote a model for joint distribution indexed by a possibly infinite-dimensional parameter α, and conditional and marginal distributions are defined analogously. We require Assumption 10.

Assumption 10. The following hold for a set of control outcomes $\mathcal{C} \subset [p]$ and for any α:

1. (Negative control outcomes) $(Y_{\mathcal{C}^c}, X) \perp\!\!\!\perp Y_{\mathcal{C}} \mid U$.
2. (Equivalence) any $\widetilde{f}(y_{\mathcal{C}}, u)$ that solves $f(y_{\mathcal{C}}; \alpha) = \int \widetilde{f}(y_{\mathcal{C}}, u; \alpha)\, \mathrm{d}u$ can be written as $\widetilde{f}(y_{\mathcal{C}}, u) = f(y_{\mathcal{C}}, v^{-1}(u); \alpha)$ for some invertible but not necessarily known function v.
3. (Completeness) for all $u \in \mathcal{U}$, $f(u) > 0$; for any square-integrable function g, $\mathbb{E}[g(U) \mid Y_{\mathcal{C}}, X = x; \alpha] = 0$ almost surely if and only if $g(U) = 0$ almost surely.

The causal diagram under Assumption 101 is given by Figure 4.4. Assumption 102 is a high-level assumption stating that at any level of covariates, the joint distribution of control outcomes and confounders lies in a class where each model is identified upon a one-to-one transformation of U. In contrast to Miao et al. [125, Assumption 2 (ii)] that concern the joint distribution of $(X, U, Y_{\mathcal{C}})$, Assumption 102 only requires equivalence on the joint distribution of $(U, Y_{\mathcal{C}})$; though we also require an extra completeness assumption on U in Assumption 103 for recovering an equivalent distribution of (X, U). The completeness property plays a pivotal role in statistics [100]. Intuitively, it precludes the degeneration of the (conditional) distributions on their supports, which guarantees the uniqueness of the solution to certain linear integral equations. At different levels of X, Assumption 102 requires that any infinitesimal variability in U is accompanied by variability in $Y_{\mathcal{C}}$, which implicitly requires the dimension of $Y_{\mathcal{C}}$ to be larger than the one of U. The completeness is viewed as a regularity condition, and more detailed discussions can be found in Miao et al. [125, Appendix 2]. Building upon the approach by Miao et al. [125], we propose a modified identification approach.

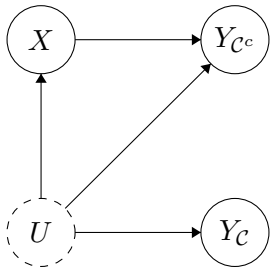

Figure 4.4: Causal diagram with negative control outcomes $Y_{\mathcal{C}}$, from which an embedding function $Y_{\mathcal{C}} \mapsto U$ can be estimated.

Theorem 21 (Nonparametric identification)**.** Under Assumptions 9 and 10, for any admissible distribution $\widetilde{f}(y_{\mathcal{C}}, u)$ that solves $f(y_{\mathcal{C}}) = \int \widetilde{f}(y_{\mathcal{C}}, u)\, \mathrm{d}u$ and let $\widetilde{f}(u) := \int \widetilde{f}(y_{\mathcal{C}}, u)\, \mathrm{d}y_{\mathcal{C}}$, there exist a unique solution $\widetilde{f}(x \mid u)$ to the equation

$$f(x) = \int \widetilde{f}(x \mid u)\widetilde{f}(u)\, \mathrm{d}u. \tag{4.2.2}$$

Let $\widetilde{f}(y_{\mathcal{C}}, u \mid x) := \widetilde{f}(y_{\mathcal{C}}, u)\widetilde{f}(x \mid u)/f(x)$, then there exists a unique solution $\widetilde{f}(y_{\mathcal{C}^c} \mid x, u)$ to the equation

$$f(y \mid x) = \int \widetilde{f}(y_{\mathcal{C}^c} \mid x, u)\widetilde{f}(y_{\mathcal{C}}, u \mid x)\, \mathrm{d}u, \tag{4.2.3}$$

Further, the potential outcome distribution is identified by

$$f_{Y(x)}(y) = \int \widetilde{f}(y_{\mathcal{C}^c} \mid u, x)\widetilde{f}(y_{\mathcal{C}}, u)\, \mathrm{d}u.$$

Theorem 21 suggests that if the joint distribution of $(Y_{\mathcal{C}}, U)$ can be estimated up to inverse transformation, then one can recover the joint distribution of potential outcome $Y(x)$. Based on Theorem 21, an operational strategy is given in two steps. The first step is to derive $\widetilde{f}(y_{\mathcal{C}}, u)$, which retrieves a proxy of U using the information from multiple control outcomes $Y_{\mathcal{C}}$. Given $\widetilde{f}(y_{\mathcal{C}}, u)$, the conditional treatment distribution $\widetilde{f}(x \mid u)$ and the condition outcome distribution can be obtained by solving integral equations (4.2.2) and (4.2.3). Even though $\widetilde{f}(y_{\mathcal{C}}, u)$ might not be unique, the estimated condition distributions $\widetilde{f}(x \mid u)$ and $\widetilde{f}(y_{\mathcal{C}^c} \mid x, u)$ are guaranteed to be unique for any given $\widetilde{f}(y_{\mathcal{C}}, u)$. Motivated by the nonparametric identification condition presented in Theorem 21, we will provide a detailed description of the deconfounding strategy for recovering the true main effect under more relaxed assumptions in the next subsection.

Remark 6 (Deconfounding with negative control outcomes)**.** The deconfounding strategy given in Theorem 21 is similar to previous negative control outcome approaches [175, 189] under parametric modeling assumptions but somewhat different from Miao et al. [125, Theorem 1] under nonparametric modeling assumptions. More specifically, Theorem 1 of Miao et al. [125] aims to recover the joint distribution of three variables (Z, X, U), where Z is an auxiliary variable that satisfies exclusion restriction condition $Z \perp\!\!\!\perp Y_{\mathcal{C}^c} \mid (X, U)$. When Z is negative control outcome $Y_{\mathcal{C}}$, we are able to factorize the joint distribution into two conditional distributions of $X \mid U$ and $Y_{\mathcal{C}} \mid U$. This property allows us to derive nonparametric identification with weaker assumptions in Theorem 21.

Another related approach is the proximal causal inference framework that uses both negative control outcomes and negative control exposures [126]. The key to their method is a bridge function $b(Y_{\mathcal{C}}, a)$ such that

$$p(Y_{\mathcal{C}^c} \mid U, A = a) = \int b(y_{\mathcal{C}}, a) p(y_{\mathcal{C}} \mid U, A = a) \, \mathrm{d}y_{\mathcal{C}} = \int b(y_{\mathcal{C}}, a) p(y_{\mathcal{C}} \mid U) \, \mathrm{d}y_{\mathcal{C}}.$$

If the bridge function b is known, then the counterfactual distributions of $Y_{\mathcal{C}^c}(a)$ can be recovered under classical causal assumptions. The proximal causal inference framework aims to bypass the estimation of the unmeasured confounders by estimating the bridge function using other extra information (e.g. negative control exposures), while our strategy relies on multiple control outcomes to estimate the distribution of confounders (up to invertible transformation) directly. With multiple negative control outcomes as in Figure 4.4, one can also split these outcomes into two nonoverlapping sets to serve the role of negative control outcomes and exposures in order to apply the proximal causal inference method; however, our approach avoids the splitting.

Remark 7 (Deconfounding with multiple treatments)**.** When there is a single outcome, and the information of confounders solely comes from multiple treatments, we can marginalize the unknown conditional distribution $f(u \mid y, x)$ over the response y to obtain $f(u \mid x) = \int f(u \mid y, x) f(y \mid x) \, \mathrm{d}y$. This suggests a two-stage procedure as in [125], for successively identifying solutions $f(u, x)$ and $f(y \mid u, x)$ from two integral equations: $f(x) = \int f(u, x) \, \mathrm{d}u$ and $f(y \mid x) = \int f(y \mid u, x) f(u \mid x) \, \mathrm{d}u$. The information used to estimate the confounders in their setting is from multiple null treatments instead of multiple outcomes. For this reason, they require strong assumptions to distinguish the set of confounded treatments associated with confounders.

4.2.2 Assumption-Lean semiparametric inference

The nonparametric identification results aim to reveal the counterfactual distributions from confounded observational data, which is useful for designing general deconfounding strategies, yet remains impractical. When restricted to semiparametric models, however, one can design more

efficient estimation and inferential procedures. A leading example of semiparametric regression models is the partially linear regression [66, 144]:

$$\mathbb{E}[Y \mid X, U] = \beta^\top X + h(U), \tag{4.2.4}$$

where Y is a high-dimensional vector of response, X is a low-dimensional vector of covariates (including the treatment of interest), $U \in \mathbb{R}^r$ is a low-dimensional latent vector, i.e., an unmeasured confounder, $\beta \in \mathbb{R}^{d\times p}$ is the coefficient to be estimated, and $h : \mathbb{R}^r \to \mathbb{R}^p$ is an unknown function. In the past decades, much attention has been on estimating and testing partially linear models.

When U is known, the coefficient β can be obtained with the double residual methodology [144], by noting that

$$\mathbb{E}[Y \mid X, U] - \mathbb{E}[Y \mid U] = \beta^\top (X - \mathbb{E}[X \mid U]),$$

More specifically, the double residual methodology proceeds in two steps: (1) regressing Y on U to obtain the residual $Y - \widehat{\mathbb{E}}[Y \mid U]$, and regress X on U to obtain the residual $X - \widehat{\mathbb{E}}[X \mid U]$; and (2) regressing the residual $Y - \widehat{\mathbb{E}}[Y \mid U]$ on the residual $X - \widehat{\mathbb{E}}[X \mid U]$. Here, the notation $\widehat{\mathbb{E}}$ denotes the estimated regression function. The resulting regression coefficient is an estimator of β. Intuitively, this procedure removes the confounding effect of U by taking the residuals, so that the final regression only captures the relationship between X and Y conditional on U, which is β under the partial linear model assumption. In the special case with binary treatments, the resulting estimator is called E-estimator [142].

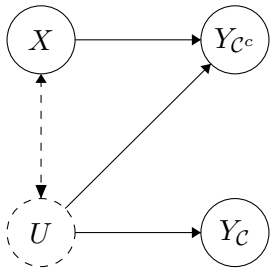

Figure 4.5: The causal relationship between X and U in Figure 4.4 can be further relaxed.

Even when the model (4.2.4) is misspecified, the estimator from the two-step procedure is directly informative about the conditional association between X and U. Under mild moment assumptions on the conditional covariance matrix of X given U, it returns a meaningful estimand

$$\begin{aligned}\beta &= \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}\mathbb{E}[\mathrm{Cov}(X, \mathbb{E}[Y \mid X, U] \mid U)] \\ &= \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}\mathbb{E}[\mathrm{Cov}(X, Y \mid U)],\end{aligned} \tag{4.2.5}$$

which itself does not crucially rely on the restrictions imposed by the outcome model (4.2.4).

Remark 8 (Relexation of causal relationship)**.** Under the causal setting in Section 4.2.1, when U is not a confounder but a moderator as in Figure 4.3, adjusting for U can also help to reduce the variance. If U is a confounder, it is necessary to adjust for U to have a proper interpretation of the main effect of X on Y. However, when U is missing, in general, we will not be certain whether U is a confounder or not. In particular, each entry of U can either be a confounder, a mediator, or a moderator (as in Figure 4.5). When targeting the estimand (4.2.5), we do not need to impose specific causal assumptions. In contrast, (4.2.5) allows us to relax the relationship between U and X, as long as the variability of X given U persists.

In summary, statistical inference targeting at projected direct effect (4.2.5) is model-free and assumption-lean. Because U is unmeasured, we rely on the strategy offered by Theorem 21 to estimate and perform inference with negative control outcomes. Our deconfounding procedure is summarized in Algorithm 3 for general link functions. Below, we describe the main steps of the procedure with an identity link as a special case.

(1) Reduction Suppose that $\mathcal{C} \subseteq [p]$ is the set of negative control outcomes such that $\beta_{\mathcal{C}} = 0$. In the first step, we aim to estimate U from the negative control outcomes $Y_{\mathcal{C}}$ independently of

Algorithm 3 Post-Integrated inference (PII) with negative control outcomes

Input: A data set $\mathcal{D}$ that contains N i.i.d. samples of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^p$, a set of control genes $\mathcal{C} \subset [p]$, and a user-specified link function g.

1: (Optional) Split sample $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ with $|\mathcal{D}_0| = m, |\mathcal{D}_1| = n$ and $N = m + n$; otherwise set $\mathcal{D} = \mathcal{D}_0 = \mathcal{D}_1$ and $N = m = n$.

2: **Estimation of the embedding functional:** Based on samples in $\mathcal{D}_0$, obtain an estimate $\widehat{f}_e : \mathbb{R}^{|\mathcal{C}|} \to \mathbb{R}^r$ for the embedding map $f_e : Y_{\mathcal{C}} \mapsto U$.

3: **Extract estimated latent embeddings:** Compute $\widehat{U} = \widehat{f}_e(Y_{\mathcal{C}})$ on $\mathcal{D}_1$.

4: **Semiparametric inference of the main effect estimand:** Use Algorithm C.5.7 to estimate

$$\widetilde{\beta}_{\cdot j} = \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U})]^{-1}\mathbb{E}[\mathrm{Cov}(X, g(\mathbb{E}[Y_j \mid X, \widehat{U}] \mid \widehat{U}))], \qquad j \in \mathcal{C}^c$$

and the empirical variance. Construct the confidence interval or compute p-values according to the asymptotic distribution of $\widetilde{\beta}$.

Output: Return the confidence intervals or p-values.

X. To distinguish from the previous causal setting, we call U as the embedding of $Y_{\mathcal{C}}$. This typically involves learning some (nonlinear) embedding map $f_e : \mathbb{R}^{|\mathcal{C}|} \to \mathbb{R}^r$ with $Y_{\mathcal{C}} \mapsto U$.

One can use the same set of data to learn the embedding function $\widehat{f}_e$ and obtain the transformed embedding $\widehat{U} = \widehat{f}_e(Y_{\mathcal{C}})$. For example, perform the principle component analysis and use the first few principal components as the estimation embedding $\widehat{U}$. In a more general scenario, we can also borrow extra datasets to estimate the embedding function. For genomic studies, many single-cell atlas of healthy cells can be used to estimate it, which helps to improve the estimation of latent embedding and is commonly used in practice for transfer learning [64].

Remark 9 (Negative control genes)**.** For genomic studies, housekeeping genes can serve as negative control outcomes. Furthermore, even though most of the genes are measured, typically only the top thousands of highly variable genes are used for the subsequent differential expression testing. It is believed that the remaining genes with low expression behave similarly under different experimental conditions. As we demonstrate later in Section 4.5, we can ideally utilize these extra genes as pseudo-negative control outcomes to improve statistical inference. Of course, there are chances that some of the genes with low expression are indeed affected by the conditions; our framework would still provide reasonable interpretability as well as robustness against such misspecification of the negative controls.

(2) Estimation In the second stage, recall that our target estimand is β in (4.2.5). Because U is unobserved, the best we can do is to use $\widehat{U}$ as the estimated embedding and focus on the estimand:

$$\widetilde{\beta}_{\cdot j} = \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U})]^{-1}\mathbb{E}[\mathrm{Cov}(X, Y_j \mid \widehat{U})], \qquad j \in \mathcal{C}^c. \tag{4.2.6}$$

This estimand quantifies the conditional associations of X and Y given $\widehat{U}$. One would typically restrict the estimation of main effects to the complement set of control genes $\mathcal{C}^c$, while for notational simplicity, we simply set $\widetilde{\beta}_{\cdot\mathcal{C}} = 0_{d\times|\mathcal{C}|}$ and present the main effect matrix $\widetilde{\beta} \in \mathbb{R}^{d\times p}$ in its whole. Note that for $j \in \mathcal{C}$, one always has $\beta_{\cdot j} = 0_d$, because $\mathbb{E}[Y_j \mid X, U] = \mathbb{E}[Y_j \mid U]$ does not depend on X and the conditional covariance between X and $\mathbb{E}[Y_j \mid X, U]$ is always zero.

(3) Inference In the last step, to provide uncertainty quantification, we rely on the efficient influence function for $\widetilde{\beta}$, similar to E-estimator [32] and two-stage least squares estimators [142, 171].

The details of semiparametric inference will be given later in Section 4.3.2 and Appendix C.5.3 for linear and nonlinear link functions, respectively.

Remark 10 (Assumption-lean and model-free inference)**.** The above procedure is minimally dependent on assumptions regarding the data-generating process. It operates independently of any underlying data model, making it truly model-free. To compute an estimate of (4.2.6), arbitrary nonparametric methods can be employed to estimate the nuisance regression function. Inference can then be performed using the efficient influence function within the semiparametric framework [171]. As we will see in the next section, this approach only requires mild moment conditions on the true regression function and consistency assumptions on the nuisance function estimation.

The procedure is straightforward and easy to understand. However, caution is warranted for nuisance regression functions and variance estimation [171]. To understand the exact conditions under which this method is effective, a more sophisticated analysis is required to quantify the bias using estimated latent embeddings. Additionally, theoretical guarantees of valid inference need to take into account the presence of multivariate covariates and multiple outcomes. The next section serves these purposes.

4.3 Statistical properties with estimated embeddings

4.3.1 Bias of main effects

Before presenting our analysis of the estimation errors, we introduce several technical assumptions. To begin with, we consider a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and let $\widehat{U}_m$ explicitly indicate the dependency of $\widehat{U}$ on $m \in \mathbb{N}$, which is the sample size used to estimate the embedding functional $\widehat{f}_e$. In general, $\widehat{U}_m$ can have different dimensions than U; to ease our theoretical analysis, we will treat the latent dimension r as known so that $\widehat{U}_m \in \mathbb{R}^r$. As we will see later, such a requirement can be weakened under certain working models. Let $\{\mathscr{F}_m\}_{m\in\mathbb{N}}$ be a filtration generated by $\{\widehat{U}_m\}_{m\in\mathbb{N}}$ such that $\mathscr{F}_m = \sigma(\widehat{U}_m)$ and $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$, and define the sub-$\sigma$-field $\mathscr{F}_\infty = \sigma(\cup_m \mathscr{F}_m) \subseteq \mathscr{F}$. We require the following assumption.

Assumption 11 (Latent embedding estimation)**.** There exists a $\mathscr{F}_\infty$-measurable and invertible function v such that $\widehat{U}_m \xrightarrow{\text{a.s.}} v(U)$. Further, $\ell_m := \|\widehat{U}_m - v(U)\|_{\mathrm{L}_2} < \infty$.

In many scenarios when we have prior information on the embedding function f_e, both the number of latent dimensions and the embedding can be consistently estimated. For example, consistent estimation of the number of latent variables has been well established under factor models [9] and under mixture models [25]. Generally, a rate of $\ell_m = \mathcal{O}_{\mathbb{P}}(m^{-\frac{1}{2}})$ can be obtained for factor analysis when there are sufficient many negative control outcomes such that $|\mathcal{C}| > m$ [8]. For mixture models, this reduces estimating the cluster membership because one can treat the one hot vector of cluster memberships as the embedding and the cluster centers as the loading, akin to factor analysis. When f_e is estimated nonparametrically by $\widehat{f}_e$, the estimated embedding $\widehat{U}_m$ can be viewed as nonparametrically generated covariates. In this context, Assumption 11 only requires the (conditional) L_2-norm of the estimation error $\widehat{f}_e - f_e$ decays to zero in probability to ensure meaningful and accurate estimation of U, which is weaker than Assumption 2 of Mammen et al. [115] that requires the (conditional) L_∞-norm of $\widehat{f}_e - f_e$ is $o_{\mathbb{P}}(1)$. Finally, we also remark that one can use extra data sources to obtain a better estimate of $\widehat{f}_e$ with a larger sample size m. In many applications, such as single-cell data analysis, the embedding function can be derived from previous alas studies so that m will be sufficiently large enough.

The following Assumption 12 imposes boundedness condition on the population quantities, and Assumption 13 imposes smoothness assumption on the regression function.

Assumption 12 (Regularity conditions)**.** There exists constants $\bar{\sigma} \geq \sigma > 0$ and $M > 0$ such that $\sigma I_d \preceq \mathbb{E}[\mathrm{Cov}(X \mid U)] \preceq \bar{\sigma} I_d$, $\sigma I_d \preceq \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U}_m)]$, $\|\beta\|_{2,\infty} \leq M$, $\|X\|_{\mathrm{L}_2} \leq M, \max_{j \in \mathcal{C}^c} \|Y_j\|_{\mathrm{L}_2} \leq M$.

Assumption 13 (Lipschitzness of regression functions)**.** The regression functions satisfy Lipschitz conditions:

$$\begin{aligned} \|\mathbb{E}[X \mid U = u_1] - \mathbb{E}[X \mid U = u_2]\| &\leq L_X \|u_1 - u_2\| \\ \|\mathbb{E}[Y_j \mid X, U = u_1] - \mathbb{E}[Y_j \mid X, U = u_2]\| &\leq L_Y \|u_1 - u_2\|, \qquad \forall\, j \in \mathcal{C}^c, \end{aligned}$$

almost surely for all $u_1, u_2 \in \mathcal{U}$ and some constants L_X and L_Y.

Assumption 13 imposes certain smoothness restrictions on the conditional expectation. In certain applications, the Lipschitz condition holds for many continuous multivariate distributions. For example, suppose W and V are jointly normally distributed with

$$\begin{pmatrix} W \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_W \\ \mu_V \end{pmatrix}, \begin{pmatrix} \Sigma_W & \Sigma_{WV} \\ \Sigma_{WV}^\top & \Sigma_V \end{pmatrix} \right).$$

Then $h(v) = \mathbb{E}[W \mid V = v] = \mu_W + \Sigma_{WV}\Sigma_V^{-1}(v - \mu_V)$ is L-Lipschitz in ℓ_2-norm, with $L = \|\Sigma_{WV}\Sigma_V^{-1}\|$. Other examples of such a regression function include the posterior mean of the exponential and Poisson distributions under their conjugate prior, as in Bayesian inference. Similar conditions have been employed for nonparametric regression with generated covariates; see, for example, Assumption 4 in Mammen et al. [115]. In particular, Mammen et al. [115] require differentiability and Lipschitz condition in ℓ_∞ of the condition expectation, which is much stronger than Assumption 13.

Consider two population coefficients β and $\widetilde{\beta}$ as defined in (4.2.5) and (4.2.6), respectively. We next quantify the difference between the two in Theorem 22.

Theorem 22 (Bias for two-stage regression with estimated covariates)**.** Under Assumptions 11–13, when $\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2} < \sigma/(2M)$, it holds that

$$\max_{j \in \mathcal{C}^c} \|\widetilde{\beta}_{\cdot j} - \beta_{\cdot j}\| \lesssim \left(\|X\|_{\mathrm{L}_2} (L_X^{\frac{1}{2}} + L_Y^{\frac{1}{2}}) + \max_{j \in \mathcal{C}^c} \|Y_j\|_{\mathrm{L}_2} L_Y^{\frac{1}{2}} \right) \ell_m.$$

Theorem 22 suggests that the upper bound of estimation error using estimated embeddings is related to the second moments of X and Y, as well as the accuracy of latent embedding estimation. This deterministic result only concerns the population quantities. Given i.i.d. samples of (X, U, Y), the corresponding estimator of $\beta_{\cdot j}$ based on finite samples is given by

$$b_{\cdot j} = (\mathbb{P}_n\{(X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2}\})^{-1} \mathbb{P}_n\{(X - \widehat{\mathbb{E}}[X \mid U])(Y_j - \widehat{\mathbb{E}}[Y_j \mid U])\}, \tag{4.3.1}$$

where $A^{\otimes 2} := AA^\top$ denotes Gram matrix of $A^\top$, and $\widehat{\mathbb{E}}[X \mid U]$ and $\widehat{\mathbb{E}}[Y \mid U]$ are the estimated nuisance functions. Because U is unobserved, we treat $\widehat{U}$ as the truth and estimate $\widetilde{\beta}_{\cdot j}$ with:

$$\widetilde{b}_{\cdot j} = (\mathbb{P}_n\{(X - \widehat{\mathbb{E}}[X \mid \widehat{U}])^{\otimes 2}\})^{-1} \mathbb{P}_n\{(X - \widehat{\mathbb{E}}[X \mid \widehat{U}])(Y_j - \widehat{\mathbb{E}}[Y_j \mid \widehat{U}])\}. \tag{4.3.2}$$

As an example, we consider a special case when the regression functions are linear models. To distinguish from previous notations, we use bold font to indicate the latent embedding matrix

$\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and its estimate $\widehat{\boldsymbol{U}} \in \mathbb{R}^{n\times \widehat{r}}$, where the latter may have a different dimension $\widehat{r}$ than the truth r. Lemma 23 below shows that we are still able to quantify the empirical estimation error of the main effects in terms of the estimation error of linear projection matrices in finite samples.

Lemma 23 (Empirical bias with estimated embeddings under linear models)**.** Define $S = \mathbb{P}_n\{(X - \mathbb{E}[X \mid U])^{\otimes 2}\}$, $\widetilde{S} = \mathbb{P}_n\{(X - \mathbb{E}[X \mid \widehat{U}])^{\otimes 2}\}$, and $\Gamma = \mathrm{diag}(\mathbb{P}_n\{YY^\top\})$. Assume S and $\widetilde{S}$ have full rank, and $\kappa(S)\|P^\perp_{\widehat{\boldsymbol{U}}} - P^\perp_{\boldsymbol{U}}\| < 1$, where for any matrix $A \in \mathbb{R}^{n\times p}$, $P_A = A(A^\top A)^{-1}A^\top$ denotes the projection matrix and $\kappa(A) = \|A\|\|A^{-1}\|$ denotes the condition number of matrix A. When $\widehat{\mathbb{E}}[X \mid U]$, $\widehat{\mathbb{E}}[X \mid \widehat{U}]$, and $\widehat{\mathbb{E}}[Y \mid U]$ are linear functions, it holds that

$$\max_{j\in\mathcal{C}^c} \|\widetilde{b}_{\cdot j} - b_{\cdot j}\| \le \left(\|b\|_{2,\infty} + \|S\|_{\mathrm{op}}^{-\frac{1}{2}}\|\Gamma\|_\infty\right) \frac{\kappa(S)\|P^\perp_{\widehat{\boldsymbol{U}}} - P^\perp_{\boldsymbol{U}}\|}{1 - \kappa(S)\|P^\perp_{\widehat{\boldsymbol{U}}} - P^\perp_{\boldsymbol{U}}\|},$$

where $\|A\|_{2,\infty} = \max_{j\in[p]} \|A_{\cdot j}\|$ is the maximum column euclidean norm for matrix $A \in \mathbb{R}^{n\times p}$.

Compared to Theorem 22, Lemma 23 suggests that the rate condition of $\widehat{\boldsymbol{U}}$ can be weakened to the rate condition of the linear projection $P^\perp_{\widehat{\boldsymbol{U}}}$. The conclusion of Lemma 23 is fully deterministic and its proof relies on the backward error analysis in numerical linear algebra [166]. The dimension of the estimated embedding is allowed to differ from the truth, as long as the column space of $\widehat{U}$ captures essential information of the column space of U. Analogously, it is possible to relax Assumption 11 to varying latent dimension settings for Theorem 22 under general data models. In this regard, one can consider a decomposition of $\lim_m \widehat{U}_m = T + A$, where T and A are a sufficient statistic and an ancillary statistic, respectively, as when U is viewed as a parameter. We leave such an extension as future work.

4.3.2 Doubly robust semiparametric inference

In the previous section, we showed that the target estimands $\widetilde{\beta}$ and β are similar whenever $\widehat{U}$ is consistent to U up to any invertible transformation. Based on the estimated embedding $\widehat{U}$, our target of estimation and inference becomes $\widetilde{\beta}$ as defined in (4.2.6). To consider potential nonparametric models for the nuisance functions, in what follows, we require the estimated nuisance functions $\widehat{\mathbb{E}}[X \mid U]$ and $\widehat{\mathbb{E}}[Y \mid U]$ to be computed from independent samples of $\mathbb{P}_n$. The required independence is very standard in recent developments of double machine learning and causal inference [84, 171], because sample splitting and cross-fitting can be used to fulfill this requirement, though one can also restrict to Donsker classes to avoid sample splitting [84].

Before we inspect the estimation error of $\widetilde{b}$ to the target estimand $\widetilde{\beta}$, we introduce one extra assumption on the moments and consistency of nuisance estimation.

Assumption 14 (Bounded moments and consistency)**.** There exists $\delta \in (0, 1]$, $M > 0$, such that

$$\|X - \mathbb{E}[X \mid \widehat{U}]\|_{\mathrm{L}_{2(1+\delta^{-1})}} \vee \|X - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_{2(1+\delta^{-1})}} \vee \|Y - \mathbb{E}[Y \mid \widehat{U}]\|_{\mathrm{L}_{2(1+\delta^{-1})}} < M,$$

$$\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_{2(1+\delta)}}, \|\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} = o_{\mathbb{P}}(1).$$

Let $O = (X, \widehat{U}, Y) \in \mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p$ denote the observation when the estimated embedding function $\widehat{f}_e$ is treated as fixed. The following theorem shows the linear expansion of the estimator $\widetilde{b}$ and gives the error bound of the residual term with high probability.

Theorem 24 (Linear expansion)**.** Consider the above inferential procedure, suppose Assumptions 12 and 14 hold and two nuisance functions $\widehat{\mathbb{E}}[X \mid \widehat{U}]$ and $\widehat{\mathbb{E}}[Y \mid \widehat{U}]$ are estimated from independent samples of $\mathbb{P}_n$. Then, the estimator $\widetilde{b}$ admits a linear expansion:

$$\sqrt{n}(\widetilde{b} - \widetilde{\beta}) = \sqrt{n}\widetilde{\Sigma}^{-1}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(O; \mathbb{P})\} + \xi,$$

where $\widetilde{\Sigma} := \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U})]$ and $\widetilde{\varphi}$ is the influence function of $\widetilde{\Sigma}\widetilde{\beta}$ defined as

$$\widetilde{\varphi}(O; \mathbb{P}) := (X - \mathbb{E}[X \mid \widehat{U}])((Y - \mathbb{E}[Y \mid X]) - \widetilde{\beta}^\top (X - \mathbb{E}[X \mid \widehat{U}]))^\top. \tag{4.3.3}$$

For any $\epsilon > 0$, there exists a constant $C = C(\epsilon, \sigma, M, L)$, such that with probability at least $1 - \epsilon$, the remainder term ξ satisfies that

$$\begin{aligned}
\|\xi\|_{2,\infty} \leq\ & C\{\|(\mathbb{P}_n - \mathbb{P})\{(X - \mathbb{E}[X \mid \widehat{U}])^{\otimes 2}\}\|_{\mathrm{op}} \\
& + \|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_{2(1+\delta)}} + \|\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_\infty\|_{\mathrm{L}_{2(1+\delta)}}\} \\
& + C\sqrt{n}\{\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2}^2 \\
& + ML\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_{\mathrm{L}_2,\infty}^2 \\
& + \|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2}\}.
\end{aligned}$$

Theorem 24 provide a non-asymptotic uniform error bound for the residual terms over multiple outcomes. With the law of large numbers and the consistency in Assumption 14, we know that the first term of the upper bound is $o_{\mathbb{P}}(1)$. On the other hand, the secondary term is also negligible under specific rate conditions on the estimation errors of nuisances. Considering an asymptotic regime when viewing m and p as sequences indexed by n and $n, m, p \to \infty$, the above result suggests the asymptotic normality, as presented in the following corollary.

Corollary 25 (Doubly robust inference with estimated emebeddings)**.** Under conditions in Theorem 24, if further, the estimation error rates of nuisance functions satisfy that $\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2}^2 = o_{\mathbb{P}}(n^{-\frac{1}{2}})$, $\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_{\mathrm{L}_2,\infty}^2 = o_{\mathbb{P}}(n^{-\frac{1}{2}})$, $\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2} = o_{\mathbb{P}}(n^{-\frac{1}{2}})$, then the estimator $\widetilde{b}$ is asymptotically normal:

$$\sqrt{n}(\widetilde{b}_{\cdot j} - \widetilde{\beta}_{\cdot j}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, \widetilde{\Sigma}^{-1}\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\widetilde{\Sigma}^{-1}), \qquad j = 1, \ldots, p.$$

Furthermore, if the conditions of Theorem 22 hold with $\ell_m = o(n^{-\frac{1}{2}})$, then we have

$$\sqrt{n}(\widetilde{b}_{\cdot j} - \beta_{\cdot j}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, \widetilde{\Sigma}^{-1}\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\widetilde{\Sigma}^{-1}), \qquad j = 1, \ldots, p.$$

In the presence of estimated embedding $\widehat{U}_m$, the influence function $\widetilde{\varphi}$ implicitly depends on the sample size m. Therefore, establishing the asymptotic normality requires verification of the Lindeberg condition for triangular array of random variables. In Corollary 25, the rate of estimation for the two nuisance functions may be slower than the parametric rate $n^{-\frac{1}{2}}$, as long as each individual estimation rate is faster than $n^{-\frac{1}{4}}$. This flexibility enables us to employ more versatile machine learning algorithms for nuisance function estimation while maintaining the validity of our inference. Furthermore, Corollary 25 suggests that efficient influence regarding the true main effect β is possible when the rate of consistently estimating the embedding is $\ell_m = o_{\mathbb{P}}(n^{-\frac{1}{2}})$. As discussed above, under factor models, one has $\ell_m = \mathcal{O}_{\mathbb{P}}(m^{-\frac{1}{2}})$, this requires

Algorithm 4 Semiparametric inference for main effects

Input: Reponses Y, covariate X, and estimated latent embedding $\widehat{U}$.

1: Use machine learning methods to obtain nuisance estimates $\widehat{\mathbb{E}}[Y \mid \widehat{U}]$ and $\widehat{\mathbb{E}}[X \mid \widehat{U}]$.

2: Fit a linear regression of $Y - \widehat{\mathbb{E}}[Y \mid \widehat{U}] \sim X - \widehat{\mathbb{E}}[X \mid \widehat{U}]$ without an intercept to obtain an estimate $\widetilde{b}$ as defined in (4.3.2) of $\widetilde{\beta}$ as defined in (4.3.2).

3: Estimate the variance of $\widetilde{b}_{\cdot j}$ by $\widehat{S}_j/n$ based on Theorem 24, where $\widehat{S}_j = \widehat{\Sigma}^{-1}\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O;\widehat{\mathbb{P}})\}\widehat{\Sigma}^{-1}$ and $\widehat{\Sigma} = \mathbb{P}_n\{(X - \widehat{\mathbb{E}}[X \mid \widehat{U}])^{\otimes 2}\}$.

Output: Confidence intervals and p-values based on asymptotic null distribution $\widetilde{b}_{\cdot j} \dot{\sim} \mathcal{N}_d(\widetilde{\beta}_{\cdot j}, \frac{\widehat{S}_j}{n})$.

$n = o(m)$, i.e., the factor loadings need to be estimated from more observations than those used for the estimation and inference of $\widetilde{b}$.

Based on Corollary 25, the data-adaptive procedure to obtain the confidence intervals and p-values is given in Algorithm 4. To fulfill the independence assumptions, one can use cross-fitting to ensure that different samples are used for step 1 and step 2. When this holds, the following proposition shows that overall Type-I error control can be controlled at the desired level. In Proposition 26, when the unit vector v is chosen to be the basis vector, it reduces to testing whether a specific covariate has zero association with individual outcomes.

Proposition 26 (Multiple linear hypothesis testing)**.** Let $t_j = \sqrt{n}\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O;\widehat{\mathbb{P}})\}^{\frac{1}{2}}\widehat{\Sigma}^{-1}\ (\widetilde{b}_{\cdot j} - \widetilde{\beta}_{\cdot j})$ be the standardized vector. For any unit vector $v \in \mathbb{R}^d$, consider the hypothesis $\mathcal{H}_{0j} : v^\top \beta_{\cdot j} = 0$. Let $\mathcal{N}_p = \{j \mid v^\top \beta_{\cdot j} = 0, j = 1, \ldots, p\}$ be the true null hypotheses. Under the assumptions of Corollary 25, as $m, n, p, |\mathcal{N}_p| \to \infty$ such that $\ell_m = o(n^{-1/2})$, it holds that $|\mathcal{N}_p|^{-1}\sum_{j\in\mathcal{N}_p} \mathbb{1}\{|v^\top t_j| > z_{\frac{\alpha}{2}}\} \xrightarrow{\mathrm{p}} \alpha$.

Remark 11 (Multiple testing)**.** The condition $\|\|\mathbb{E}[Y \mid \widehat{U}] - \widehat{\mathbb{E}}[Y \mid \widehat{U}]\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} = o_{\mathbb{P}}(1)$ in Assumption 14 controls the envelope of the regression function estimation errors. This is useful when the number of outcomes p grows in the number of sample size n, when multiple testing procedures based on multiplier bootstrap can be applied to control both the family-wise error rate and the false discovery rate [49]. Alternatively, one can simply apply the Benjamini–Hochberg procedure for multiple testing corrections.

4.4 Simulation

We generate the data from generalized partial linear models. The covariate $X \in \mathbb{R}$ is sampled from $\mathcal{N}(0, 1)$; the latent variable $U = X\alpha + \epsilon \in \mathbb{R}^r$ is a linear function of X, where $r = 10$, $\alpha_{1j} \sim \mathrm{Unif}(-1, 1)$ and $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2)$ independently for $j \in [r]$; and the response is generated from generalized linear models with a Logistic link $\mathrm{logit}(\mathbb{E}[Y \mid X, U]) = X\beta + U\eta$, where $\beta_{1j} \sim 2 \times \mathrm{Bernoulli}(0.2)$ and $\sqrt{r} \cdot \eta_{ij} \sim \mathrm{Unif}(-1, 1)$ independently for $i \in [r]$ and $j \in [p]$. We set the total number of outcomes to be $p = 1000$, and use 500 null outcomes as the negative outcomes.

We evaluate four methods: (1) Glm (X): naive generalized linear models that use Logistic regression that only uses observed covariate X to predict Y; (2) Glm (X, U): oracle Logistic regression that uses both observed covariate X and latent variable U to predict Y; (3) Pii (X, U): the proposed post-integrated inference method that uses observed covariate X and latent embedding U to predict Y; and (4) Pii $(X, \widehat{U})$: the proposal method that uses the first r PCs of the outcome matrix are selected as $\widehat{U}$.

For Pii, we use the random forest to estimate the nuisance functions $\mathbb{E}[X \mid U]$, $\mathbb{E}[Y \mid X, U]$,

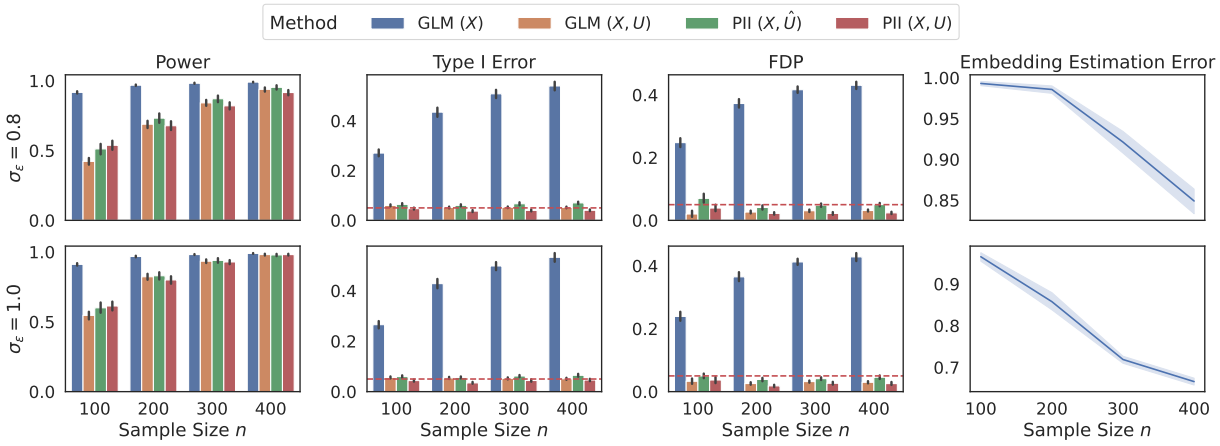

Figure 4.6: Simulation results with 500 negative control outcomes out of a total of 1000 outcomes. For PII, the nuisance functions are estimated using random forests. The data model is the Logistic regression model. The first and second rows have noise levels $\sigma_\epsilon = 0.8$ and $\sigma_\epsilon = 1$, respectively, for the latent variables.

and $\mathbb{E}[g(\mathbb{E}[Y \mid X, U]) \mid U]$ and apply extrapolated cross-validation (ECV) [44] to select the hyperparameter that minimizes the estimated mean squared error. ECV allows us to use a smaller number of trees for estimating the out-of-sample prediction errors based on out-of-bag observations and extrapolate the risk estimation up to a larger number of trees consistently without sample splitting. In our experiment, we use 25 trees to perform ECV and the hyperparameters we consider include: 'max_depth' in $\{1, 3, 5\}$ for the depth of each tree, 'max_samples' in $\{0.25, 0.5, 0.75, 1\}$ for bootstrap samples and the number of trees in $\{1, \ldots, 50\}$.

To compare the performance of different methods, the power, type-I error, and false discovery proportion (FDP) for hypothesis testing are analyzed. For both the type-I error and power, we set the significance level to be 0.05. For FDP, we use the Benjamini-Hochberg procedure with FDR controlled at 0.05. As shown in the first two columns of Figure 4.6, the GLM-NAIVE regression method fails to control the inflated type-I error, resulting in numerous false positives. Furthermore, as the sample size increases, this method becomes even more anti-conservative. Conversely, the GLM-ORACLE regression method exhibits tight control over type-I error, as expected. When the latent embedding U is known, we observe that PII also effectively controls type-I error. Additionally, under certain conditions, PII provides greater power than the GLM-ORACLE. This may be attributed to PII's ability to address the effect of collinearity between X and U on the nonlinear outcome models through a two-step procedure, whereas GLM-ORACLE does not, leading to conservative results.

When the latent embedding U is unknown, we evaluate the performance of the estimated $\widehat{U}$. As shown in the third panel of Figure 4.6, the error of embedding projection matrix $\|P_{\widehat{U}} - P_U\|_{\text{op}}$ decreases rapidly as the sample size n increases. When U can be well approximated, PII experiences a slightly inflated type-I error because it targets the modified main effect $\widetilde{\beta}$ instead of the true effect β. However, the statistical error remains reasonable, the FDP is controlled at the desired level, and PII achieves greater power compared to the oracle GLM in many cases. Lastly, PII exhibits greater power when the conditional variation of X given U is large (i.e., $\mathbb{V}(\epsilon)$ is relatively larger than the linear projected signal strength $\|\gamma\|$). One could potentially use the ratio of these two quantities as a metric to quantify the level of confounding.

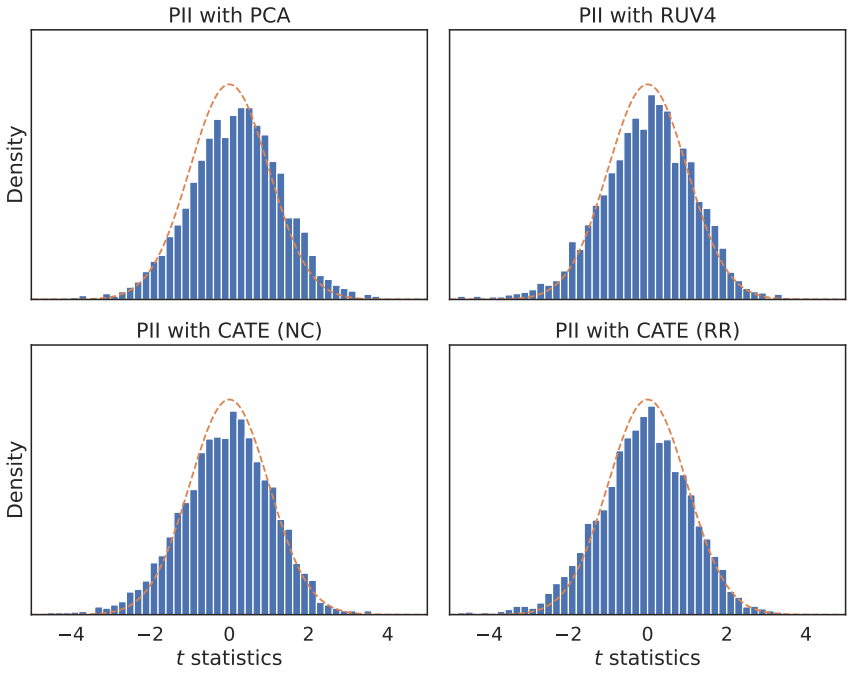

Figure 4.7: Histogram of t-statistics of *PTEN* perturbation by different methods. PCA with 50 components, RUV4, CATE-NC, and CATE-RR.

4.5 Application on single-cell CRISPR data analysis

Background In a recent study by Lalli et al. [93], the molecular mechanisms of genes associated with neurodevelopmental disorders, particularly Autism Spectrum Disorder (ASD), were investigated using a modified CRISPR-Cas9 system. Experiments focused on 13 ASD-linked gene knockdowns in Lund Human Mesencephalic neural progenitor cells, with gene expression changes assessed through single-cell RNA sequencing. The progression of neuronal differentiation was estimated via a pseudotime trajectory (Figure C.62), revealing that some genetic perturbations impact this progression.

Confounders significantly affect the interpretation of single-cell CRISPR perturbation results, as these experiments often resemble observational studies. Variables like cell size, cycle stage, and microenvironment heterogeneity can alter gene expression patterns, obscuring true genetic effects. To address these challenges, we use 4000 lowly variable genes as negative control outcomes for adjustment, focusing on 4163 highly variable genes for differential expression analysis on 8320 cells. The data preprocessing procedure is detailed in Appendix C.6.2.

Compared methods and embedding estimation. We compare the proposed method with four methods for hypothesis testing: (1) GLM: Score tests based on generalized linear models with Negative Binomial likelihood and log link function. The covariance matrix is estimated using the HC3-type robust estimator. This method does not adjust for potential confounding effects. (2) RUV4: A heuristic method proposed by Gagnon-Bartsch and Speed [56] that uses principle components on the residual matrix of regressing the negative control outcomes on the covariate of interest to estimate the latent embeddings. Based on heuristic calculations, the authors claim that the RUV-4 estimator has approximately the oracle variance. (3) CATE-NC:

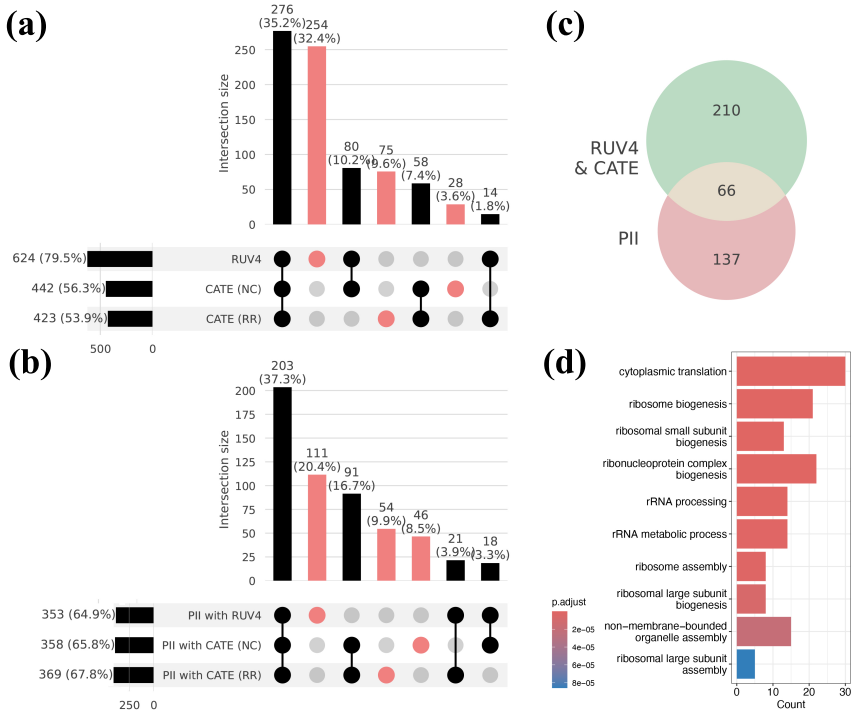

Figure 4.8: Summary of significant genes (p-values< 0.05) associated with *PTEN* perturbation by different confounder adjustment methods. (a) Upset plot of discoveries by three methods: RUV4, CATE-NC, and CATE-RR, as in Figure C.64. (b) Upset plot of discoveries by PII with embedding estimated by three methods: RUV4, CATE-NC, and CATE-RR, as in Figure 4.7. (c) The Venn plot of two sets of discoveries. One set includes 276 common discoveries by RUV4, CATE-NC, and CATE-RR, while the other includes 203 common discoveries by PII with the same estimated embeddings given by the three methods. (d) Gene ontology analysis of 137 distinct discoveries by PII.

The deconfounding method CATE proposed by Wang et al. [175] with negative controls, which uses maximum likelihood estimation to estimate the latent embedding. Under simplified Gaussian linear models, they show that their estimator has asymptotical type I error control when the number of negative controls is large. (4) CATE-RR: A variant of CATE method [175] with robust regression, which is similar to the heuristic algorithm LEAPP [160] and utilizes the sparsity of effects to estimate the latent embeddings.

For PII, we use four methods to estimate the cell embeddings, including PCA, RUV4, CATE-NC, and CATE-RR. The first three methods use negative control to estimate the embedding, while the last is only valid under the sparsity assumption on the effects. Before running PCA, we follow the preprocessing procedure in single-cell data analysis to adjust the library size of each cell to be 10^4, add one pseudo count, and take the logarithm. We then select the top 50 principal components as the estimated embeddings. For the last three embedding estimation methods, we supply all 13086 genes as input, specify the set of pseudo-negative control genes when applicable, and set the number of factors to 10, a value commonly used by researchers based on empirical evidence. Though not presented in the paper, we observed similar results with higher numbers of factors.

Results. The study by Lalli et al. [93] indicates that some perturbations affect changes in gene expressions along pseudotime, potentially altering development speed. Using pseudotime as a covariate allows us to examine if perturbations explain effects beyond developmental changes; see Appendix C.6.2 for the extended background of the data. Biologically, we expect more signals on pseudotime states (Figure C.63) than on perturbation conditions.

We focus on the target gene *PTEN*, which is crucial in neural development and differentiation and influences other genes in a cascading manner when it is knockdown [93]. Examining the empirical distribution of test statistics for perturbation conditions using GLM reveals conservative results for genes like *CTNND2*, *MECP2*, and *MYT1L* (Figure C.64). This suggests that GLM without adjusting for hidden confounders leads to biased hypothesis testing. PII corrects these biases (Figure 4.2, Figure 4.7), and even PCA-based simple embedding estimation effectively calibrates test distributions. Comparing methods RUV4, CATE-NC, CATE-RR with PII, we see PII reduces distinct discoveries from 45.6% to 38.8% (Figure 4.8(a)), indicating more coherent outcomes with PII.

To further assess the biological significance of the discoveries, we examine 276 and 203 common discoveries (Figure 4.8(a) and Figure 4.8(b)). Significant discrepancies are noted (Figure 4.8(c)). The associated gene ontology terms on biological processes using `clusterProfiler` package with default false discovery control threshold [181] reveal that unlike the 210 genes from RUV4, CATE-NC, and CATE-RR, which had no associated GO terms, the 137 genes unique to PII align with ribosome-related processes (Figure 4.8(d)), supporting studies on *PTEN*'s impact on these processes [34, 103].

Chapter 5

Discussion

In Chapter 2, we presented novel estimation and inference procedures for multivariate generalized linear models with unmeasured confounders in the high-dimensional scenarios when both the sample size n and response size p tend to infinity. Our approach consists of three main steps. In the first step, we disentangle the marginal effects from the uncorrelated confounding effects, recovering the column space of latent coefficients $\widehat{\boldsymbol{\Gamma}}$ from the latter. We provide non-asymptotic estimation error bounds for both the estimated natural parameter matrix $\widehat{\boldsymbol{\Theta}}_0$ and the projection onto the column space of $\widehat{\boldsymbol{\Gamma}}$. In the second step, we estimate both latent factors $\boldsymbol{Z}$ and primary effects $\boldsymbol{B}$ by solving a constrained lasso-type problem that confines $\boldsymbol{B}$ to the orthogonal space of $\widehat{\boldsymbol{\Gamma}}$. From the column-wise estimation error of the latent components, we obtain the estimation error for the primary effects in the presence of nuisance parameters. In the third step, we design an inferential procedure to correct the bias introduced by ℓ_1-regularization and establish Type-I error and family-wise error rate controls. Numerically, we demonstrate the usage of the proposed method with Poisson and Negative Binomial likelihoods for bulk-cell and single-cell simulations, respectively. Compared to alternative methods, the proposed method effectively controls the Type-I error and false discovery proportion while delivering enhanced statistical power and precision as the count data get sparser and more over-dispersed. Furthermore, our analysis of real single-cell datasets underscores the importance of accounting for confounding effects when major covariates are unobserved. Notably, our proposed method consistently outperforms alternative techniques, demonstrating superior precision and specificity, thus establishing its suitability for high-dimensional sparse count data.

The present study, while offering valuable insights, is not without its limitations and opportunities for future exploration. Some of these include the development of hypothesis testing for confounding effects, the theoretical guarantee of the FDR, and more robust criteria for selecting the optimal number of latent factors. Recent works by Dai et al. [35] and Chen and Li [26] offer promising insights that may contribute to resolving some of these challenges. Although we have briefly touched upon the applicability of our proposed method under non-canonical link functions in Appendix A.6.4, comprehensive theoretical guarantees remain an area deserving of further research and investigation.

In Chapter 3, we inspect more flexible semiparametric approaches. The rapid growth of high-throughput single-cell technologies has created an urgent need for robust causal inference frameworks capable of disentangling treatment effects from confounding influences. Existing methods, such as CINEMA-OT [38], have advanced the field by separating confounder and treatment signals and providing per-cell treatment-effect estimates. However, these methods rely on

the assumption of no unmeasured confounders, which is often violated in observational studies and in vivo experiments. Additionally, many confounder adjustment methods, such as RUV [140], depend on linear model assumptions that do not directly model count data or provide robust differential expression testing at the gene level. Addressing these limitations, causarray introduces a doubly robust framework that integrates generalized confounder adjustment with semiparametric inference to enable reliable and interpretable causal analysis.

causarray directly models count data using generalized linear models for unmeasured confounder estimation, overcoming a key limitation of RUV in DE analysis. Unlike CINEMA-OT [38] and CoCoA-diff [134], which rely on optimal transport or matching techniques, causarray employs a doubly robust framework that combines flexible machine learning models with semiparametric inference. This approach enhances stability and interpretability while enabling valid statistical inference of treatment effects. In an in vivo Perturb-seq study of ASD/ND genes, causarray uncovered gene-level perturbation effects that were missed by prior module-based analyses. It identified biologically relevant pathways linked to neuronal development and synaptic functions for multiple autism risk genes. Similarly, in a case-control study of Alzheimer's disease using three human brain transcriptomic datasets, causarray revealed consistent causal gene expression changes across datasets and highlighted key biological processes such as synaptic signaling and cell development. These findings underscore the ability of causarray to provide biologically meaningful insights across diverse contexts.

Despite its strengths, causarray has certain limitations. Its performance depends on the accurate estimation of unmeasured confounders, which may vary with dataset complexity and experimental design. Furthermore, while causarray provides robust DE testing, its integration with advanced spatial or trajectory analysis frameworks remains unexplored [43, 188]. Future research could focus on extending causarray to incorporate prior biological knowledge or extrapolate to unseen perturbation-cell pairs, similar to emerging methods like CPA [108]. Such advancements would further enhance its applicability in single-cell causal inference on general omics.

In Chapter 4, a potential concern of the proposed method is whether the estimated embeddings might act as colliders, especially if $\widehat{U}$ is influenced by both X and $Y_{\mathcal{C}^c}$. However, our fundamental assumption is that $Y_{\mathcal{C}}$ is driven by a low-dimensional embedding U but not the covariate X, which inherently mitigates the risk of $\widehat{U}$ becoming a collider. If this foundational assumption does not hold, the direct effect estimand (4.2.5) might not align with researchers' interests, necessitating the use of domain knowledge to identify and investigate alternative target estimands. Despite design-free deconfounding with negative control outcomes, other strategies exist (detailed comparison in Appendix C.2). While our framework allows for flexible machine learning algorithms, it introduces computational complexity, especially with increasing outcomes and hyperparameter tuning. For practical applications, specialized models like variational autoencoders for joint outcome function fitting [39, 127] and efficient cross-validation methods can be beneficial.

Further extensions involve incorporating interaction effects [171], developing tests for nonparametric confounding [124]. Explorations into settings with high-dimensional latent embeddings and covariates [125, 183] could also be of interest. In real data analysis, we use pseudo-negative control outcomes, which can be viewed as one variant of the synthetic control approaches [1]. Providing theoretical guarantees for the valid construction of negative control outcomes and selective importance features [46] in the presence of correlations from data remains an area of practical interest.

Appendix A

Simultaneous inference for generalized linear models with unmeasured confounders

A.1 Proof of Proposition 1

Proof of Proposition 1. Because the one-parameter exponential family is minimal, the natural parameter space is convex, and the log-partition function A is strictly convex. Based on the information of the first moment of $\boldsymbol{y}$ and the log-partition function A, we can identify $\boldsymbol{B}\boldsymbol{x}+\boldsymbol{\Gamma}\boldsymbol{z} = \boldsymbol{\theta} = A'^{-1}(\mathbb{E}[\boldsymbol{y}])$. Because $\boldsymbol{\Gamma}\boldsymbol{w}$ has zero mean and is uncorrelated to $\boldsymbol{x}$, $\mathrm{Cov}(\boldsymbol{\Gamma}\boldsymbol{w}) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ can be identified as the residual covariance of regression of $\boldsymbol{\theta}$ on $\boldsymbol{x}$.

Because $\lambda_r(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top) \geq \tau_p$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}_w$ have full rank. Let $\boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{U}_r^\top$ be the reduced eigenvalue decomposition of $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ where $\boldsymbol{U}_r \in \mathbb{R}^{p\times r}$. Note that

$$
\begin{aligned}
\mathcal{P}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2}(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top \\
&= \boldsymbol{U}_r\boldsymbol{\Lambda}_r^{\frac{1}{2}}(\boldsymbol{\Lambda}_r^{\frac{1}{2}}\boldsymbol{U}_r^\top\boldsymbol{U}_r\boldsymbol{\Lambda}_r^{\frac{1}{2}})^{-1}\boldsymbol{\Lambda}_r^{\frac{1}{2}}\boldsymbol{U}_r^\top \\
&= \boldsymbol{U}_r\boldsymbol{\Lambda}_r^{\frac{1}{2}}(\boldsymbol{\Lambda}_r^{\frac{1}{2}}\boldsymbol{\Lambda}_r^{\frac{1}{2}})^{-1}\boldsymbol{\Lambda}_r^{\frac{1}{2}}\boldsymbol{U}_r^\top \\
&= \boldsymbol{U}_r\boldsymbol{U}_r^\top.
\end{aligned}
$$

Thus, $\mathcal{P}_{\boldsymbol{\Gamma}}$ can be recovered.

By the orthogonal decomposition, we have $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}\boldsymbol{B} + \mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}$. Let $\boldsymbol{e}_{p,i} = (\delta_{i\ell})_{1\leq\ell\leq p}$ and

$\boldsymbol{e}_{d,j} = (\delta_{j\ell})_{1\le\ell\le d}$. We consider the (i,j)-th entry of $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}$:

$$\begin{aligned}
|\boldsymbol{e}_{p,i}^\top \mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}\boldsymbol{e}_{d,j}| &= |\boldsymbol{e}_{p,i}^\top \boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2}(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{B}\boldsymbol{e}_j| \\
&\le \|\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2}(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,i}\|_\infty \cdot \|\boldsymbol{B}\boldsymbol{e}_{d,j}\|_1 \\
&= \max_{\ell\in[p]} |\boldsymbol{e}_\ell^\top \boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2}(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,i}| \cdot \|\boldsymbol{B}\boldsymbol{e}_{d,j}\|_1 \\
&\le \max_{\ell\in[p]} \|\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,\ell}\|_2 \cdot \|(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,i}\|_2 \cdot \|\boldsymbol{B}\boldsymbol{e}_{d,j}\|_1 \\
&\le \max_{\ell\in[p]} \|\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,\ell}\|_2 \cdot \|(\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2})^{-1}\|_{\mathrm{op}} \cdot \|\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,i}\|_2 \cdot \|\boldsymbol{B}_j\|_1 \\
&\le \max_{\ell\in[p]} \|\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,\ell}\|_2 \cdot \lambda_r(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top)^{-1} \cdot \|\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{e}_{p,i}\|_2 \cdot \|\boldsymbol{B}_j\|_1 \\
&= \mathcal{O}\left(\frac{\|\boldsymbol{B}_j\|_1}{\tau_p}\right) \\
&= o(1), \qquad \text{(A.1.1)}
\end{aligned}$$

where the first two inequalities are from Holder's inequality; the third inequality holds because of the sub-multiplicativity of the operator norm; and the last inequality holds because $\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w^{1/2}$ and $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_w\boldsymbol{\Gamma}^\top$ have the same non-zero eigenvalues. Then we have

$$\|\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B}\|_{\mathrm{F}} \le \sqrt{p}\max_{1\le i\le p}\|\boldsymbol{B}^\top\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{e}_{p,i}\|_2 \lesssim \frac{\sqrt{p}\|\boldsymbol{B}\|_{1,1}}{\tau_p}.$$

Thus, the conclusion follows. □

A.2 Estimation error of natural parameters by alternative maximization

In this section, we gather useful results to bound the estimation error for the natural parameter matrix. Let $E_C = \{\boldsymbol{\Theta}^* \in \mathcal{R}_C^{n\times p}\}$ be the event that all the natural parameters are bounded. From Assumption 1, we know that $\mathbb{P}(E_C) = \iota_n \to 1$ as $n \to \infty$. Under event E_C, because A is strictly convex and trice continuously differentiable, we have that

$$\kappa_1 := \inf_{\theta\in\mathcal{R}_C} A''(\theta) > 0 \text{ and } \kappa_2 := \sup_{\theta\in\mathcal{R}_C} A''(\theta) < \infty. \qquad \text{(A.2.1)}$$

These facts enable us to derive Theorem 2, which will be used in Appendix A.3 for proving Theorem 3 and in Appendix A.4 for proving Theorem 5.

A.2.1 Estimation error of natural parameters

Proof of Theorem 2. We split the proof into two parts under event E_C.

Part (1) Bounding $\|\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$. From the assumption of Theorem 2, we have

$$\mathcal{L}(\boldsymbol{\Theta}^*) - \mathcal{L}(\widehat{\boldsymbol{\Theta}}_0) \ge 0,$$

which also holds when $\widehat{\mathbf{\Theta}}_0$ is the maximum likelihood estimator. From Lemma A.2.1 it further follows that

$$0 \le \sqrt{2(d+r)}\|\boldsymbol{Y} - A'(\mathbf{\Theta}^*)\|_{\mathrm{op}}\|\widehat{\mathbf{\Theta}}_0 - \mathbf{\Theta}^*\|_{\mathrm{F}} - \frac{\kappa_2}{2}\|\widehat{\mathbf{\Theta}}_0 - \mathbf{\Theta}^*\|_{\mathrm{F}}^2.$$

Thus, we have

$$\|\widehat{\mathbf{\Theta}}_0 - \mathbf{\Theta}^*\|_{\mathrm{F}} \le \frac{2\sqrt{2(d+r)}}{\kappa_2}\|\boldsymbol{Y} - A'(\mathbf{\Theta}^*)\|_{\mathrm{op}}.$$

Next, we bound the operator norm of $\boldsymbol{Y} - A'(\mathbf{\Theta}^*)$. Conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}^*$, observe that $e_{ij} := y_{ij} - A'(\theta^*_{ij})$ ($i \in [n]$ and $j \in [p]$) are independent, zero-mean, and sub-exponential with parameters $\nu = \sqrt{\kappa_2}$ and $\alpha = 1/C^2$. To see this, note that its moment generating function is $\mathbb{E}[\exp(te_{ij})] = \exp(A(\theta^*_{ij} + t) - A(\theta^*_{ij}) - tA'(\theta^*_{ij})) = \exp(A''(\theta^*_{ij} + t')t^2/2)$ for some $|t'| < |t|$. By Assumption 1, we have $\mathbb{E}[\exp(te_{ij})] \le \kappa_2 t^2/2$ for all $|t| < C^2$, which shows that e_{ij} is sub-exponential. By Lemma A.2.2, for any $\delta > 0$, with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta}$, it follows that

$$\begin{aligned}\|\widehat{\mathbf{\Theta}}_0 - \mathbf{\Theta}^*\|_{\mathrm{F}} &\le \frac{2\sqrt{2(d+r)}}{\kappa_2}(4\nu\sqrt{n \vee p} + 2\delta^{\frac{3}{2}}\sqrt{c}(\alpha \vee \nu)\log(np)\sqrt{\log(n+p)}) \\ &\lesssim \sqrt{(d+r)(n \vee p)}.\end{aligned}$$

Part (2) Bounding $\max_{1\le j\le p}\|(\widehat{\mathbf{\Theta}}_0)_j - \mathbf{\Theta}^*_j\|_2$. Similarly to Part (1), by union bound, we have

$$\begin{aligned}\max_{1\le j\le p}\|(\widehat{\mathbf{\Theta}}_0)_j - \mathbf{\Theta}^*_j\|_2 &\le \max_{1\le j\le p}\frac{2\sqrt{2(d+r)}}{\kappa_2}\|\boldsymbol{Y}_j - A'(\mathbf{\Theta}^*_j)\|_2 \\ &\le \frac{2\sqrt{2(d+r)}}{\kappa_2}(4\nu\sqrt{n} + 2(\delta+1)^{\frac{3}{2}}\sqrt{c}(\alpha \vee \nu)\log(n)\sqrt{\log(n+1)}),\end{aligned}$$

with probability at least $1 - p(n+p)^{-\delta-1} - p(np)^{-\delta-1} \ge 1 - (n+p)^{-\delta} - (np)^{-\delta}$, for any $\delta > 0$.

For $\delta > 1$, taking union bound over the above two events and E_C finishes the proof. □

A.2.2 Technical lemmas

Lemma A.2.1 (Upper bound of likelihood difference)**.** Suppose that $\mathbf{\Theta}_1 \in \mathcal{R}_{r_1}, \mathbf{\Theta}_2 \in \mathcal{R}_{r_2}$ with $r_j = \mathrm{rank}(\mathbf{\Theta}_j)$ for $j = 1, 2$. Define $\kappa_1 := \inf_{\theta\in\mathcal{R}} A''(\theta)$. Then it holds that

$$\mathcal{L}(\mathbf{\Theta}_2) - \mathcal{L}(\mathbf{\Theta}_1) \le \frac{\sqrt{r_1 + r_2}}{n}\|\boldsymbol{Y} - A'(\mathbf{\Theta}_2)\|_{\mathrm{op}}\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\mathrm{F}} - \frac{\kappa_1}{2n}\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\mathrm{F}}^2,$$

and

$$\mathcal{L}(\mathbf{\Theta}_2) - \mathcal{L}(\mathbf{\Theta}_1) \le \frac{\sqrt{r_1 + r_2}}{n}\|\boldsymbol{Y} - A'(\mathbf{\Theta}_1)\|_{\mathrm{op}}\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\mathrm{F}} + \frac{\kappa_2}{2n}\|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|_{\mathrm{F}}^2.$$

Proof of Lemma A.2.1. Recall that $\mathcal{L}(\mathbf{\Theta}) = n^{-1}[-\mathrm{tr}(\boldsymbol{Y}^\top\mathbf{\Theta}) + \mathrm{tr}(\mathbf{1}_{p\times n}A(\mathbf{\Theta}))]$. Then we have

$$\begin{aligned}\mathcal{L}(\mathbf{\Theta}_2) - \mathcal{L}(\mathbf{\Theta}_1) = &\frac{1}{n}\mathrm{tr}((\boldsymbol{Y} - A'(\mathbf{\Theta}_2))^\top(\mathbf{\Theta}_1 - \mathbf{\Theta}_2)) \\ &- \frac{1}{n}\mathrm{tr}(\mathbf{1}_{p\times n}(A(\mathbf{\Theta}_1) - A(\mathbf{\Theta}_2)) - A'(\mathbf{\Theta}_2)^\top(\mathbf{\Theta}_1 - \mathbf{\Theta}_2)).\end{aligned} \tag{A.2.2}$$

Next, we analyze the two terms separately.

For the first term, we have

$$\begin{aligned} &\operatorname{tr}((\boldsymbol{Y}-A'(\boldsymbol{\Theta}_2))^\top(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)) && \text{(A.2.3)}\\ \le&\sqrt{\operatorname{rank}(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)}\|\boldsymbol{Y}-A'(\boldsymbol{\Theta}_2)\|_{\mathrm{op}}\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|_{\mathrm{F}}\\ \le&\sqrt{\operatorname{rank}(\boldsymbol{\Theta}_1)+\operatorname{rank}(\boldsymbol{\Theta}_2)}\|\boldsymbol{Y}-A'(\boldsymbol{\Theta}_2)\|_{\mathrm{op}}\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|_{\mathrm{F}}, && \text{(A.2.4)}\end{aligned}$$

where the first inequality is from the matrix norm inequality $|\operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{B})| \le \sqrt{\operatorname{rank}(\boldsymbol{B})}\|\boldsymbol{A}\|_{\mathrm{op}}\|\boldsymbol{B}\|_{\mathrm{F}}$ and the last inequality is due to the fact that $\operatorname{rank}(\boldsymbol{A}+\boldsymbol{B}) \le \operatorname{rank}(\boldsymbol{A})+\operatorname{rank}(\boldsymbol{B})$.

For the second term, note that each entry inside the trace takes the form

$$\begin{aligned} A((\boldsymbol{\Theta}_1)_{ij})-A((\boldsymbol{\Theta}_2)_{ij})-A'((\boldsymbol{\Theta}_2)_{ij})((\boldsymbol{\Theta}_1)_{ij}-(\boldsymbol{\Theta}_2)_{ij}) &= \frac{1}{2}A''(\theta)((\boldsymbol{\Theta}_1)_{ij}-(\boldsymbol{\Theta}_2)_{ij})^2\\ &\ge \frac{\kappa_1}{2}((\boldsymbol{\Theta}_1)_{ij}-(\boldsymbol{\Theta}_2)_{ij})^2\end{aligned}$$

where the first equality is from Taylor expansion with θ lies between $(\boldsymbol{\Theta}_1)_{ij}$ and $(\boldsymbol{\Theta}_2)_{ij}$ and the definition of $\kappa_1 := \inf_{\theta\in\mathcal{R}} A''(\theta) \ge 0$. Thus we have

$$\operatorname{tr}(\mathbf{1}_{p\times n}(A(\boldsymbol{\Theta}_1)-A(\boldsymbol{\Theta}_2))-A'(\boldsymbol{\Theta}_2)^\top(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)) \ge \frac{\kappa_1}{2}\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|_{\mathrm{F}}^2. \qquad \text{(A.2.5)}$$

Combining (A.2.2), (A.2.4), and (A.2.5) finishes the proof of the first inequality. Similarly, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}_2)-\mathcal{L}(\boldsymbol{\Theta}_1) &= \frac{1}{n}\operatorname{tr}((\boldsymbol{Y}-A'(\boldsymbol{\Theta}_1))^\top(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2))\\ &\quad+\frac{1}{n}\operatorname{tr}(\mathbf{1}_{p\times n}(A(\boldsymbol{\Theta}_2)-A(\boldsymbol{\Theta}_1))-A'(\boldsymbol{\Theta}_1)^\top(\boldsymbol{\Theta}_2-\boldsymbol{\Theta}_1))\\ &\le \frac{\sqrt{r_1+r_2}}{n}\|\boldsymbol{Y}-A'(\boldsymbol{\Theta}_1)\|_{\mathrm{op}}\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|_{\mathrm{F}}+\frac{\kappa_2}{2n}\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|_{\mathrm{F}}^2,\end{aligned}$$

which completes the proof of the second inequality. □

Lemma A.2.2 (Operator norm of matrices with sub-exponential entries)**.** Let $\boldsymbol{X} = (x_{ij})_{i\in[n],j\in[p]}$ be a matrix with independent and centered entries such that x_{ij}'s are (ν,α)-sub-exponential random variables[1] with parameters $\nu,\alpha>0$:

$$\mathbb{E}[\exp(tx_{ij})] \le \exp(t^2\nu^2/2), \qquad \forall\ |t|<\frac{1}{\alpha}.$$

Then for all $\delta>0$, there exists a universal constant $c>0$ such that, with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}$, when n,p are large enough, it holds that

$$\|\boldsymbol{X}\|_{\mathrm{op}} \le 4\nu\sqrt{n\vee p}+2\delta^{3/2}\sqrt{c}(\alpha\vee\nu)\log(np)\sqrt{\log(n+p)}.$$

Proof of Lemma A.2.2. We define a symmetric matrix

$$\boldsymbol{Z} = (z_{ij}) = \begin{pmatrix} 0 & \widetilde{\boldsymbol{X}} \\ \widetilde{\boldsymbol{X}}^\top & 0\end{pmatrix} \in \mathbb{R}^{(n+p)\times(n+p)},$$

[1]Here we adopt the definition from Wainwright [174, Definition 2.7]. In some literature, the term 'sub-gamma' is used interchangeably with 'sub-exponential' to refer to this definition.

where $\widetilde{\boldsymbol{X}} := (\widetilde{x}_{ij}) = \boldsymbol{X} - \boldsymbol{X}'$ and $\boldsymbol{X}' = \left(x'_{ij}\right)$ is an independent copy of $\boldsymbol{X}$. Because $\widetilde{x}_{ij}$'s have symmetric distribution and are independent, it follows that z_{ij}'s are also independent and symmetric random variables, and $\|\boldsymbol{Z}\|_{\mathrm{op}} = \|\widetilde{\boldsymbol{X}}\|_{\mathrm{op}}$. Because the tail event $\mathbb{1}\{\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{op}} \geq t\}$ is a convex function on $\boldsymbol{X}'$, by Jensen's inequality we have

$$\mathbb{1}\{\|\boldsymbol{X}\|_{\mathrm{op}} \geq t\} = \mathbb{1}\{\|\boldsymbol{X} - \mathbb{E}_{\boldsymbol{X}'}[\boldsymbol{X}']\|_{\mathrm{op}} \geq t\} \leq \mathbb{E}_{\boldsymbol{X}'}[\mathbb{1}\{\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{op}} \geq t\}] = \mathbb{E}_{\boldsymbol{X}'}[\mathbb{1}\{\|\widetilde{\boldsymbol{X}}\|_{\mathrm{op}} \geq t\}].$$

By Fubini's theorem, it follows that

$$\mathbb{P}(\|\boldsymbol{X}\|_{\mathrm{op}} \geq t) \leq \mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\mathbb{1}\{\|\widetilde{\boldsymbol{X}}\|_{\mathrm{op}} \geq t\}] = \mathbb{P}(\|\widetilde{\boldsymbol{X}}\|_{\mathrm{op}} \geq t) = \mathbb{P}(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq t).$$

Then, it suffices to bound the tail probability of $\|\boldsymbol{Z}\|_{\mathrm{op}}$.

We define a truncated random matrix $\boldsymbol{Z}(\lambda)$ of $\boldsymbol{Z}$,

$$\boldsymbol{Z}(\lambda) = (z_{ij}(\lambda))_{1\leq i\leq n, 1\leq j\leq p} = (z_{ij}\,\mathbb{1}\left(|z_{ij}| \leq \lambda\right))_{1\leq i\leq n, 1\leq j\leq p}$$

whose entries are independent, symmetric random variables bounded by λ. By Bandeira and van Handel [11, Corollary 3.12], there exists a universal constant $c > 0$ such that

$$\mathbb{P}\left(\|\boldsymbol{Z}(\lambda)\|_{\mathrm{op}} \geq 2^{\frac{3}{2}} \max_{1\leq i\leq n+p}\left(\sum_{j=1}^{n+p}\mathbb{E}[z_{ij}^2(\lambda)]\right)^{\frac{1}{2}} + t\right) \leq (n+p)\exp\left(-\frac{t^2}{c\lambda^2}\right).$$

Note that

$$\begin{aligned}
\max_{1\leq i\leq n+p}\left(\sum_{j=1}^{n+p}\mathbb{E}[z_{ij}^2(\lambda)]\right)^{\frac{1}{2}} &\leq \max_{1\leq i\leq n+p}\left(\sum_{j=1}^{n+p}\mathbb{E}[z_{ij}^2]\right)^{1/2} \\
&= \max\left\{\max_{1\leq i\leq n}\left(\sum_{j=1}^{p}\mathbb{E}[\widetilde{x}_{ij}^2]\right)^{\frac{1}{2}}, \max_{1\leq j\leq p}\left(\sum_{i=1}^{N}\mathbb{E}[\widetilde{x}_{ij}^2]\right)^{\frac{1}{2}}\right] \\
&\leq \max\{\sqrt{p}, \sqrt{n}\}\max_{i,j}\mathbb{E}[\widetilde{x}_{ij}^2]^{\frac{1}{2}} \\
&\leq \sqrt{2(n\vee p)}\max_{i,j}\mathbb{E}[x_{ij}^2]^{\frac{1}{2}} \\
&\leq \nu\sqrt{2(n\vee p)},
\end{aligned}$$

where the first inequality is from the definition of the truncated variable, the second and the third inequality is by Cauchy-Schwartz inequality, and the last inequality is because x_{ij} is (ν, α)-sub-exponential. Thus, the above two inequality yields that

$$\mathbb{P}\left(\|\boldsymbol{Z}(\lambda)\|_{\mathrm{op}} \geq 4\nu\sqrt{n\vee p} + t\right) \leq (n+p)\exp\left(-\frac{t^2}{c\lambda^2}\right).$$

Then we have

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq 4\nu\sqrt{n\vee p} + t\right) &\leq \mathbb{P}\left(\|\boldsymbol{Z}(\lambda)\|_2 \geq 4\nu\sqrt{n\vee p} + t\right) + \mathbb{P}\left(\max_{1\leq i,j\leq n+p}|z_{ij}| > \lambda\right) \\
&\leq (n+p)\exp\left(-\frac{t^2}{c\lambda^2}\right) + \sum_{1\leq i\leq n}\sum_{1\leq j\leq p}\mathbb{P}(|\widetilde{x}_{ij}| > \lambda) \\
&\leq (n+p)\exp\left(-\frac{t^2}{c\lambda^2}\right) + np\left(\exp\left(-\frac{\lambda^2}{4\nu^2}\right)\vee\exp\left(-\frac{\lambda}{2\alpha}\right)\right).
\end{aligned}$$

where the last inequality follows because $\widetilde{x}_{ij} = x_{ij} - x'_{ij}$ is $(2\nu^2, \alpha)$-sub-exponential. For all $\delta > 0$, let $\lambda = 2(\delta+1)(\alpha \vee \nu)\log(np)$ and $np \geq 3$, the second term is bounded by $(np)^{-\delta}$. Let $t = \lambda((\delta+1)c\log(n+p))^{1/2}$, the first term is bounded by $(n+p)^{-\delta}$. Combining these yields that

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq 4\nu\sqrt{n \vee p} + 2\delta^{3/2}\sqrt{c}(\alpha \vee \nu)\log(np)\sqrt{\log(n+p)}\right) \leq (n+p)^{-\delta} + (np)^{-\delta},$$

which completes the proof. □

A.3 Estimation of latent coefficients

A.3.1 Preparatory definitions

Recall that $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{\Gamma}}$ are derived from the SVD of $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0$ from the first-stage optimization and we have $\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0$. Analogous to Bing et al. [20], we define $\boldsymbol{H}_0 := (np)^{-1}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}^{-3/2}$ and

$$\widetilde{\boldsymbol{\Gamma}} := \boldsymbol{\Gamma}^*\boldsymbol{H}_0 = (np)^{-1}\boldsymbol{\Gamma}^*\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}^{-3/2}, \tag{A.3.1}$$

which is identifiable because it depends on both the data $\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}^{-3/2}$ and the identifiable quantity $\boldsymbol{\Gamma}^*\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}$. Note that

$$\begin{aligned}
\mathcal{P}_{\widetilde{\boldsymbol{\Gamma}}} &= \widetilde{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1}\widetilde{\boldsymbol{\Gamma}}^\top \\
&= \boldsymbol{\Gamma}^*(\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\boldsymbol{V})(\boldsymbol{V}^\top\boldsymbol{\Gamma}^*\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\boldsymbol{\Gamma}^*\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\boldsymbol{V})^{-1}(\boldsymbol{V}^\top\boldsymbol{\Gamma}^*\boldsymbol{W}^{*\top}\boldsymbol{W}^*)\boldsymbol{\Gamma}^{*\top} \\
&= \boldsymbol{\Gamma}^*(\boldsymbol{\Gamma}^{*\top}\boldsymbol{\Gamma}^*)^{-1}\boldsymbol{\Gamma}^{*\top} \\
&= \mathcal{P}_{\boldsymbol{\Gamma}^*}
\end{aligned} \tag{A.3.2}$$

because both $\boldsymbol{\Gamma}^{*\top}\boldsymbol{\Gamma}^*$ and $\boldsymbol{W}^{*\top}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\boldsymbol{V} \in \mathbb{R}^{r\times r}$ have full rank. Thus, to quantify the error between $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}$ and $\mathcal{P}_{\boldsymbol{\Gamma}^*}$, we can first analyze the error between $\widehat{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Gamma}}$.

A.3.2 Proof of Theorem 3

Proof of Theorem 3. We split the proof into three parts by bounding the operator norm, column-wise ℓ_2-norm, and the sup norm consecutively.

Part (1) Bounding the operator norm. From (A.3.2), we have that $\mathcal{P}_{\widetilde{\boldsymbol{\Gamma}}} = \mathcal{P}_{\boldsymbol{\Gamma}^*}$ for $\widetilde{\boldsymbol{\Gamma}}$ defined in (A.3.1). Then we have

$$\begin{aligned}
&\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\boldsymbol{\Gamma}^*}\|_{\mathrm{op}} \\
=&\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\widetilde{\boldsymbol{\Gamma}}}\|_{\mathrm{op}} \\
=&\|\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^\top\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^\top - \widetilde{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1}\widetilde{\boldsymbol{\Gamma}}^\top\|_{\mathrm{op}} \\
\leq&\|(\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}})(\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1}\widetilde{\boldsymbol{\Gamma}}^\top\|_{\mathrm{op}} + \|\widehat{\boldsymbol{\Gamma}}((\widehat{\boldsymbol{\Gamma}}^\top\widehat{\boldsymbol{\Gamma}})^{-1} - (\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1})\widetilde{\boldsymbol{\Gamma}}^\top\|_{\mathrm{op}} + \|\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^\top\widehat{\boldsymbol{\Gamma}})^{-1}(\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}})^\top\|_{\mathrm{op}},
\end{aligned}$$

where the last inequality is from the triangle inequality. Next, we bound the three terms separately. Recall $\widetilde{\boldsymbol{\Gamma}}$ is defined in (A.3.1) with $\|\widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{op}} \asymp \|\widehat{\boldsymbol{\Gamma}}\|_{\mathrm{op}} \asymp \sqrt{p}$ by Assumption 3 and Lemma A.3.2. Then by Lemma A.3.1 and Assumption 3, for all $\delta > 0$, the first term in the above display is bounded

$$\|(\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}})(\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1}\widetilde{\boldsymbol{\Gamma}}^\top\|_{\mathrm{op}} \leq \|\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{op}}\|(\widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{\Gamma}})^{-1}\|_{\mathrm{op}}\|\widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{op}} \leq C'(n \wedge p)^{-\frac{1}{2}},$$

with probability at least $1-2(n+p)^{-\delta}-2(np)^{-\delta}-\exp(-n)$, for some constant $C'>0$ and $\delta>0$. Similarly, the third term is also bounded by $\mathcal{O}_{\mathbb{P}}\left((n\vee p)^{-1/2}\right)$. It remains to bound the second term:

$$
\begin{aligned}
&\|\widehat{\mathbf{\Gamma}}((\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}-(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1})\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}\\
=&\|\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}\\
\le&\|\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}\|_{\mathrm{op}}\cdot\|(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}\\
\lesssim&\frac{1}{\sqrt{p}}\|(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}\\
\le&\frac{1}{\sqrt{p}}\|\widehat{\mathbf{\Gamma}}^\top(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}+\frac{1}{\sqrt{p}}\|(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})^\top\widetilde{\mathbf{\Gamma}}(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}\\
\le&\frac{1}{\sqrt{p}}\|\widehat{\mathbf{\Gamma}}\|_{\mathrm{op}}\|(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\|_{\mathrm{op}}+\frac{1}{\sqrt{p}}\|\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}}\|_{\mathrm{op}}\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}\|_{\mathrm{op}}\\
\lesssim&\mathcal{O}\left((n\wedge p)^{-\frac{1}{2}}\right),
\end{aligned}
$$

with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$. The proof for the operator norm is completed by combing the above inequality.

Part (2) Bounding the column-wise ℓ_2-norm. Let $\boldsymbol{e}_j\in\mathbb{R}^p$ be the unit vector such that its i-th entry is one if $i=j$ and zero otherwise. Similar to Part (1), note that

$$
\begin{aligned}
&\|(\mathcal{P}_{\widehat{\mathbf{\Gamma}}}-\mathcal{P}_{\mathbf{\Gamma}^*})\boldsymbol{e}_j\|_2\\
\le&\|(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2+\|\widehat{\mathbf{\Gamma}}((\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}-(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1})\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2+\|\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})^\top\boldsymbol{e}_j\|_2.
\end{aligned}
$$

The first term can be bounded analogously as

$$
\max_{1\le j\le p}\|(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2\le\|\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}}\|_{\mathrm{op}}\|(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\|_{\mathrm{op}}\max_{1\le j\le p}\|\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2\le C'[p(n\wedge p)]^{-\frac{1}{2}},
$$

for some constant $C'>0$, by noting that $\|\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2=\mathcal{O}_{\mathbb{P}}(1)$. The rest of the terms follow a similar argument as in Part (1), under the same probabilistic events therein.

Part (3) Bounding the sup norm. The sup norm $\|\cdot\|_{\max}$ can be upper bounded analogously:

$$
\begin{aligned}
&\|\mathcal{P}_{\widehat{\mathbf{\Gamma}}}-\mathcal{P}_{\mathbf{\Gamma}^*}\|_{\max}\\
=&\max_{i,j\in[p]}|\boldsymbol{e}_i^\top(\mathcal{P}_{\widehat{\mathbf{\Gamma}}}-\mathcal{P}_{\mathbf{\Gamma}^*})\boldsymbol{e}_j|\\
\le&\max_{i,j\in[p]}(|\boldsymbol{e}_i^\top(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j|+|\boldsymbol{e}_i^\top\widehat{\mathbf{\Gamma}}((\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}-(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1})\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j|+|\boldsymbol{e}_i^\top\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^\top\widehat{\mathbf{\Gamma}})^{-1}(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})^\top\boldsymbol{e}_j|)
\end{aligned}
$$

The first term can be bounded as

$$
\max_{1\le j\le p}|\boldsymbol{e}_i^\top(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j|\le\|(\widehat{\mathbf{\Gamma}}-\widetilde{\mathbf{\Gamma}})\boldsymbol{e}_i\|_{\mathrm{op}}\|(\widetilde{\mathbf{\Gamma}}^\top\widetilde{\mathbf{\Gamma}})^{-1}\|_{\mathrm{op}}\max_{1\le j\le p}\|\widetilde{\mathbf{\Gamma}}^\top\boldsymbol{e}_j\|_2\le C'[p^2(n\wedge p)]^{-\frac{1}{2}},
$$

by involving both Part (1) and (2). The rest of the terms follow a similar argument as in Part (1), under the same probabilistic events therein. This completes the proof. □

A.3.3 Technical lemmas

Lemma A.3.1 (Estimation error of $\widehat{\boldsymbol{\Gamma}}$)**.** Under Assumptions 1–3 and event E_C, for all $\delta > 0$ and sufficiently large n and p, there exists a absolute constant $C' > 0$ such that

$$\max_{1 \le j \le p} \|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j\|_2 \le C', \qquad \|\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{op}} \le C' \sqrt{\frac{n \vee p}{n}},$$

with probability at least $1 - 2(n+p)^{-\delta} - 2(np)^{-\delta} - \exp(-n)$.

Proof of Lemma A.3.1. Define $\boldsymbol{E} = \boldsymbol{W}^* \boldsymbol{\Gamma}^{*\top}$ and $\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{W}} \widehat{\boldsymbol{\Gamma}}^{\top}$[2]. Then we have $\widehat{\boldsymbol{E}} = \boldsymbol{W}^* \boldsymbol{\Gamma}^{*\top} + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = (\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*) - (\boldsymbol{X}\widehat{\boldsymbol{F}}^{\top} - \boldsymbol{X}\boldsymbol{F}^{*\top})$. By the definition of $\widehat{\boldsymbol{E}}$ we have

$$\frac{1}{np} \widehat{\boldsymbol{E}}^{\top} \widehat{\boldsymbol{E}} = \boldsymbol{V} \boldsymbol{\Sigma}^2 \boldsymbol{V}^{\top}.$$

Note that $\widehat{\boldsymbol{\Gamma}} = \sqrt{p} \boldsymbol{V} \boldsymbol{\Sigma}^{1/2}$, we further have

$$\frac{1}{np} \widehat{\boldsymbol{E}}^{\top} \widehat{\boldsymbol{E}} \widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}} \boldsymbol{\Sigma}^2$$

and

$$\frac{1}{n\sqrt{p}} \widehat{\boldsymbol{E}}^{\top} \widehat{\boldsymbol{E}} \boldsymbol{V} \boldsymbol{\Sigma}^{-3/2} = \widehat{\boldsymbol{\Gamma}}.$$

It follows that

$$\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}} = \frac{1}{n\sqrt{p}} (\boldsymbol{E}^{\top} \boldsymbol{\Delta} + \boldsymbol{\Delta}^{\top} \boldsymbol{E} + \boldsymbol{\Delta}^{\top} \boldsymbol{\Delta}) \boldsymbol{V} \boldsymbol{\Sigma}^{-3/2}.$$

Because the operator norm is sub-multiplicative, the ℓ_2-norm of the jth row of $\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}}$ is bounded by

$$\begin{aligned} \|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j\|_2 &\le \frac{1}{n\sqrt{p}} (\|\boldsymbol{E}^{\top} \boldsymbol{\Delta}_j\|_2 + \|\boldsymbol{\Delta}^{\top} \boldsymbol{E}_j\|_2 + \|\boldsymbol{\Delta}^{\top} \boldsymbol{\Delta}_j\|_2) \|\boldsymbol{V}\|_{\mathrm{op}} \|\boldsymbol{\Sigma}^{-3/2}\|_{\mathrm{op}} \\ &\le \frac{1}{n\sqrt{p}} (\|\boldsymbol{E}\|_{\mathrm{op}} \|\boldsymbol{\Delta}_j\|_2 + \|\boldsymbol{\Delta}\|_{\mathrm{op}} \|\boldsymbol{E}_j\|_2 + \|\boldsymbol{\Delta}\|_{\mathrm{op}} \|\boldsymbol{\Delta}_j\|_2) \|\boldsymbol{\Sigma}^{-3/2}\|_{\mathrm{op}}, \end{aligned} \tag{A.3.3}$$

and

$$\|\widehat{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{op}} \le \frac{1}{n\sqrt{p}} (2\|\boldsymbol{E}\|_{\mathrm{op}} \|\boldsymbol{\Delta}\|_{\mathrm{op}} + \|\boldsymbol{\Delta}\|_{\mathrm{op}}^2) \|\boldsymbol{\Sigma}^{-3/2}\|_{\mathrm{op}}. \tag{A.3.4}$$

To proceed, we split the proof into three parts.

[2]Throughout the manuscript, the notation $\boldsymbol{e}_i$ is reserved for the unit vector and is not the i-th row of $\boldsymbol{E}$. We will only use the notations of $\boldsymbol{E}$ and $\boldsymbol{E}_j$ to denote the matrix of latent components and its j-th column.

Part (1) Bounding operator noms of $\boldsymbol{\Sigma}$, $\boldsymbol{E}$, and $\boldsymbol{\Delta}$. Note that $\boldsymbol{E} = \boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ are mean-zero sub-Gaussian random vectors from Assumption 3. From Lemma A.3.2, for any $\delta > 0$, there exists $C_\Sigma > 0$, such that

$$\|\boldsymbol{E}\|_{\mathrm{op}} \le 2C_\Sigma\sqrt{np}, \qquad \frac{1}{C_\Sigma} \le \lambda_r(\boldsymbol{\Sigma}) \le \lambda_1(\boldsymbol{\Sigma}) \le C_\Sigma,$$

with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \exp(-n)$.

Because

$$\boldsymbol{\Delta} = \boldsymbol{E} - \widehat{\boldsymbol{E}} = \mathcal{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0) + \mathcal{P}_{\boldsymbol{X}}\boldsymbol{E},$$

we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{op}} &= \|\mathcal{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0) + \mathcal{P}_{\boldsymbol{X}}\boldsymbol{E}\|_{\mathrm{op}} \\ &\le \|\mathcal{P}_{\boldsymbol{X}}^{\perp}\|_{\mathrm{op}}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|_{\mathrm{op}} + \|\mathcal{P}_{\boldsymbol{X}}\boldsymbol{E}\|_{\mathrm{op}} \\ &\le \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|_{\mathrm{op}} + \|\mathcal{P}_{\boldsymbol{X}}\boldsymbol{E}\|_{\mathrm{op}}.\end{aligned}$$

On the one hand, from Theorem 2, it follows that when n, p are large enough,

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|_{\mathrm{op}} \le \sqrt{c(d+r)(n \vee p)}$$

with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta}$ for some constant $c > 0$. On the other hand, notice that

$$\mathcal{P}_{\boldsymbol{X}}\boldsymbol{E} = \mathcal{P}_{\boldsymbol{X}}\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top} = \boldsymbol{X}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{W}^*}{n}\boldsymbol{\Gamma}^{*\top}$$

where $\boldsymbol{X}^\top\boldsymbol{W}^* = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{w}_i^{*\top}$ is the sum of n i.i.d. sub-exponential random matrices with zero means. By the matrix Bernstein's inequality, $\|\boldsymbol{X}^\top\boldsymbol{W}^*/n\|_{\mathrm{op}} \lesssim \sqrt{\log(nd)/n}$. Thus, the second term can be bounded as

$$\|\mathcal{P}_{\boldsymbol{X}}\boldsymbol{E}\|_{\mathrm{op}} \lesssim \sqrt{n} \cdot 1 \cdot \sqrt{\frac{\log(nd)}{n}} \cdot \sqrt{p}$$

with probability at least $1 - n^{-\delta}$. The above results suggest that

$$\|\boldsymbol{\Delta}\|_{\mathrm{op}} \le \sqrt{c(d+r)(n \vee p)}$$

Below, we condition on the two events above, which hold with probability at least $1 - 2(n + p)^{-\delta} - 2(np)^{-\delta} - \exp(-n)$ by union bound.

Part (2) Bounding $\|\boldsymbol{E}_j\|_2$ and $\|\boldsymbol{\Delta}_j\|_2$ From Lemma A.3.3 and Assumption 3, we have

$$\max_{1\le j\le p}\|\boldsymbol{E}_j\|_2 = \max_{1\le j\le p}\|\boldsymbol{W}^*\boldsymbol{\gamma}_j^*\|_2 \le \|\boldsymbol{W}\|_{\mathrm{op}}\max_{1\le j\le p}\|\boldsymbol{\gamma}_j^*\|_2 \le 2C_\Sigma\sqrt{n}.$$

On the other hand, because $\widehat{\boldsymbol{W}}^\top\widehat{\boldsymbol{W}} = n\boldsymbol{\Sigma}$ and $\|\widehat{\boldsymbol{\gamma}}_j\|_2 \le C^2C_\Sigma^{1/2}$ for all $j \in [p]$ from Lemma A.3.4, we also have

$$\max_{1\le j\le p}\|\widehat{\boldsymbol{E}}_j\|_2 = \max_{1\le j\le p}\|\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\gamma}}_j\|_2 = \max_{1\le j\le p}\sqrt{n}\|\widehat{\boldsymbol{\gamma}}_j\|_2 \lesssim C^2C_\Sigma^{3/2}\sqrt{n}.$$

Thus, by triangle inequality, we have

$$\max_{1\le j\le p}\|\boldsymbol{\Delta}_j\|_2 \le \max_{1\le j\le p}(\|\boldsymbol{E}_j\|_2 + \|\widehat{\boldsymbol{E}}_j\|_2) \lesssim (2 + C^2)C_\Sigma\sqrt{n}.$$

Part (3) Combining the previous results. From (A.3.3), (A.3.4) and the previous two parts, we have

$$\begin{aligned}
&\max_{1\le j\le p}\|\widehat{\boldsymbol{\gamma}}_j-\widetilde{\boldsymbol{\gamma}}_j\|_2\\
\lesssim&\frac{1}{n\sqrt{p}C_\Sigma}(2C_\Sigma\sqrt{np}(2+C^2)C_\Sigma\sqrt{n}+\sqrt{c(d+r)(n\vee p)}C\sqrt{n}+\sqrt{c(d+r)(n\vee p)}(2C_\Sigma+C)\sqrt{n})\\
\lesssim&2(2+C^2)C_\Sigma
\end{aligned}$$

and

$$\|\widehat{\boldsymbol{\Gamma}}-\widetilde{\boldsymbol{\Gamma}}\|_2\le\frac{1}{n\sqrt{p}C_\Sigma}(4C_\Sigma\sqrt{np}\sqrt{c(d+r)(n\vee p)}+c(d+r)(n\vee p))\lesssim 4\sqrt{c},$$

which finishes the proof. □

Lemma A.3.2 (Spectrum of $\boldsymbol{\Sigma}$)**.** Under Assumptions 1–3 and event E_C, for all $\delta>0$ and sufficiently large n and p, there exists a absolute constant $C_\Sigma>1$ such that

$$\frac{1}{C_\Sigma}\le\lambda_r(\boldsymbol{\Sigma})\le\lambda_1(\boldsymbol{\Sigma})\le C_\Sigma,$$

with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$.

Proof of Lemma A.3.2. By Weyl's inequality, we have that for all $k\in[r]$,

$$\begin{aligned}
\left|\lambda_k(\boldsymbol{\Sigma})-\frac{1}{\sqrt{np}}\lambda_k(\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top})\right|&=\frac{1}{\sqrt{np}}\left|\lambda_k(\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Gamma}}^\top)-\lambda_k(\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top})\right|\\
&\le\frac{1}{\sqrt{np}}\|\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Gamma}}^\top-\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top}\|_{\mathrm{op}}\\
&=\frac{1}{\sqrt{np}}\|\mathcal{P}_{\boldsymbol{X}}^{\perp}(\widehat{\boldsymbol{\Theta}}_0-\boldsymbol{\Theta}^*)\|_{\mathrm{op}}\\
&\le\frac{1}{\sqrt{np}}\|\widehat{\boldsymbol{\Theta}}_0-\boldsymbol{\Theta}^*\|_{\mathrm{op}}.
\end{aligned}\tag{A.3.5}$$

We next bound $\lambda_k(\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top})$ and $\|\widehat{\boldsymbol{\Theta}}_0-\boldsymbol{\Theta}^*\|_{\mathrm{op}}$ separately. Applying Lemma A.3.3 under Assumption 3 yields that

$$\frac{1}{C_0'}\le\lambda_r\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le\lambda_1\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le C_0'.$$

with probablity at least $1-\exp(-n)$ for some constant $C_0'>1$. From Assumption 3 we further have

$$\sqrt{\frac{1}{CC_0'}}\le\frac{1}{\sqrt{np}}\lambda_r(\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top})\le\frac{1}{\sqrt{np}}\lambda_1(\boldsymbol{W}^*\boldsymbol{\Gamma}^{*\top})\le\sqrt{CC_0'}.\tag{A.3.6}$$

On the other hand, from Theorem 2 we have for all $\delta>0$, there exists $C>0$ such that

$$\frac{1}{\sqrt{np}}\|\widehat{\boldsymbol{\Theta}}_0-\boldsymbol{\Theta}^*\|_{\mathrm{op}}\le\frac{1}{\sqrt{np}}\|\widehat{\boldsymbol{\Theta}}_0-\boldsymbol{\Theta}^*\|_{\mathrm{F}}\le C\sqrt{\frac{r(n\vee p)}{np}}=:C_{n,p},\tag{A.3.7}$$

with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}$.

Condition on the above two events, applying triangle inequality on (A.3.5) and combining (A.3.6) and (A.3.7), we have

$$\sqrt{\frac{1}{CC_0'}}-C_{n,p}\le\lambda_r(\mathbf{\Sigma})\le\lambda_1(\mathbf{\Sigma})\le\sqrt{CC_0'}+C_{n,p}.$$

Note that $C_{n,p}=o(1)$ as both n and p tend to infinity. When n and p is such that $C_{n,p}<1/2\sqrt{CC_0'}$, setting $C_\Sigma=3\sqrt{CC_0'}/2$ gives the desired bound. By union bound, this holds with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$, which finishes the proof. □

Lemma A.3.3 (Spectrum of $\boldsymbol{W}^*$)**.** Under Assumption 3, for sufficiently large n, there exists a absolute constant $C_0'>0$ such that

$$\frac{1}{C_0'}\le\lambda_r\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le\lambda_1\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le C_0',$$

with probability at least $1-\exp(-n)$.

Proof of Lemma A.3.3. Note that $\boldsymbol{W}^*\in\mathbb{R}^{n\times r}$ is a random matrix whose rows $\boldsymbol{w}_1^*,\dots,\boldsymbol{w}_n^*$ are i.i.d. sub-Gaussian random vectors. From the concentration inequality of the operator norm of the random matrices [172, Theorem 4.6.1, Exercise 4.7.3], for any $u>0$ we have

$$\left\|\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*-\mathbf{\Sigma}_w\right\|_{\text{op}}\le CK^2\left(\sqrt{\frac{r+u}{n}}+\frac{r+u}{n}\right)\|\mathbf{\Sigma}_w\|_{\text{op}},$$

with probability at least $1-2\exp(-u)$, where $C>0$ is some absolute constant and $K=\max_{i\in[n]}\|\boldsymbol{w}_i^*\|_{\psi_2}$. When n is sufficiently large such that $2CK^2n^{-1/4}<1$, setting $u=n^{1/2}$ and $C'=2CK^2n^{-1/4}C_0$ yields that

$$\left\|\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*-\mathbf{\Sigma}_w\right\|_{\text{op}}\le CK^2\left(n^{-\frac{1}{4}}+n^{-\frac{1}{2}}\right)\|\mathbf{\Sigma}_w\|_{\text{op}}\le C'<C_0.$$

By Weyl's inequality, we have that,

$$\max_{k\in[r]}\left|\lambda_k\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)-\lambda_k\left(\mathbf{\Sigma}_w\right)\right|\le\left\|\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*-\mathbf{\Sigma}_w\right\|_{\text{op}}\le C'.$$

By triangle inequality and the boundedness of $\mathbf{\Sigma}_w$'s spectrum from Assumption 3, it follows that

$$\frac{1}{C_0}-C'\le\min_{k\in[r]}\lambda_k\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le\max_{k\in[r]}\lambda_k\left(\frac{1}{n}\boldsymbol{W}^{*\top}\boldsymbol{W}^*\right)\le C_0+C',$$

with probability at least $1-\exp(-n)$. Setting $C_0'=\min\{C_0+C',(C_0^{-1}-C')^{-1}\}$ finishes the proof. □

Lemma A.3.4 (Boundedness of latent factors and loadings)**.** Under Assumptions 1–3 and event E_C, for all $\delta>0$ and sufficiently large n and p, there exists a absolute constant $C_\Sigma>1$ such that $\|\widehat{\boldsymbol{\gamma}}\|_2\le C^2C_\Sigma^{1/2}$, with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$.

Proof of Lemma A.3.4. Recall that $\widehat{\boldsymbol{W}}_0$ and $\widehat{\boldsymbol{\Gamma}}_0$ are the solutions to the alternative maximization problems which satisfy that $\|\widehat{\boldsymbol{w}}_{0,i}\|_2 \le C$ and $\|\widehat{\boldsymbol{\gamma}}_{0,j}\|_2 \le C$ for $i = 1, \ldots, n$ and $1, \ldots, p$. Let $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top = \sqrt{np}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the condensed SVD. Then $\widehat{\boldsymbol{\Gamma}}$ is defined to be $\sqrt{p}\boldsymbol{V}\boldsymbol{\Sigma}^{1/2} = \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{W}}_0^\top\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}/\sqrt{n}$. From Lemma A.3.2, with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \exp(-n)$, it holds that

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}_j\|_2 &\le \|\widehat{\boldsymbol{W}}_0/\sqrt{n}\|_{\mathrm{op}}\|\boldsymbol{U}\|_{\mathrm{op}}\|\boldsymbol{\Sigma}^{1/2}\|_{\mathrm{op}}\|\widehat{\boldsymbol{\gamma}}_{0,j}\|_2 \\ &\le \max_{1\le i\le n}\|\widehat{\boldsymbol{W}}_{0,i}\|_2 \cdot 1 \cdot C_\Sigma^{1/2} \cdot C \\ &\le C_\Sigma^{1/2}C^2,\end{aligned}$$

which finishes the proof. □

A.4 Estimation of latent factors and direct effects

A.4.1 Preparatory definitions

Towards proving Theorem 5, we first introduce the following notations. Recall that optimization problem (2.3.2) is the multivariate lasso with nuisance parameter. Define the response vector $\widetilde{\boldsymbol{y}} = \mathrm{vec}(\boldsymbol{Y}) \in \mathbb{R}^{np}$, the design matrices $\widetilde{\boldsymbol{X}} = (\boldsymbol{I}_p \otimes \boldsymbol{X}) \in \mathbb{R}^{np\times pd}$, $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B}^*)$, $\boldsymbol{\zeta} = \mathrm{vec}(\boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top})$, and the projection matrix $\widetilde{\mathcal{P}}_{\boldsymbol{\Gamma}^*} = (\mathcal{P}_{\boldsymbol{\Gamma}^*} \otimes \boldsymbol{I}_d) \in \mathbb{R}^{pd\times pd}$. Here, symbol '$\otimes$' denotes the Kronecker product. With slight abuse of notations, we use $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\zeta})$ to denote the unregularized loss function of (2.3.2). Let $\mathcal{F} = \{(\boldsymbol{\beta}, \boldsymbol{\zeta}) \in \mathbb{R}^{pd} \times \mathbb{R}^{np} \mid \boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B}), \boldsymbol{\zeta} = \mathrm{vec}(\boldsymbol{Z}\boldsymbol{\Gamma}^\top)$ for $\boldsymbol{B} \in \mathbb{R}^{p\times d}, \boldsymbol{Z} \in \mathbb{R}^{n\times r}, \boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$ such that $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{B} = \boldsymbol{0}, \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\zeta} \in \mathcal{R}_C^{np}\}$ be the feasible set of $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$. Then the joint optimization problem (2.3.2) is equivalent to

$$\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}} \in \operatorname*{argmin}_{(\boldsymbol{\beta},\boldsymbol{\zeta})\in\mathcal{F}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\zeta}) + \lambda\|\boldsymbol{\beta}\|_1. \tag{A.4.1}$$

Let $(\widetilde{\boldsymbol{\beta}}^*, \widetilde{\boldsymbol{\zeta}}^*) = (\mathrm{vec}(\mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp\boldsymbol{B}^*), \mathrm{vec}(\boldsymbol{X}\boldsymbol{B}^*\mathcal{P}_{\boldsymbol{\Gamma}^*} + \boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}))$, $(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) = (\mathrm{vec}(\boldsymbol{B}^*), \mathrm{vec}(\boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}))$ denote the tuples of target coefficients and note that $(\widetilde{\boldsymbol{\beta}}^*, \widetilde{\boldsymbol{\zeta}}^*) \in \mathcal{F}$.

The organization of the following subsections is summarized as below:

- Appendix A.4.2 proves Corollary 4, which controls the column-wise ℓ_2-norm of the estimation error of the latent component.
- Appendix A.4.3 proves Theorem 5. The proof of Theorem 5 consists of three main steps:
  (1) Establish cone condition: We involve Lemma A.4.1 to show that the estimation from the sequential optimization problems (2.3.4)-(2.3.5) obtain approximately optimality condition to the joint optimization problem (2.3.2):
  $$\mathcal{L}(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{Z}}) + \lambda\|\widehat{\boldsymbol{B}}\|_{1,1} \le \mathcal{L}(\boldsymbol{B}^*, \boldsymbol{\Gamma}^*, \boldsymbol{Z}^*) + \lambda\|\boldsymbol{B}^*\|_{1,1} + \tau_{n,p},$$
  for some small order term $\tau_{n,p}$ with high probability. This enables us to derive the cone condition.
  (2) Obtain upper and lower bound of the first-order approximation error: We involve Lemma A.4.2 to derive the upper bound, and Lemma A.4.3 to establish the locally strong convexity and hence the lower bound.
  (3) Derive the estimation errors: We compute the ℓ_2-norm and ℓ_1-norm estimation error based on the previous two steps.
- Appendix A.4.4 gathers helper lemmas used in the current section.

A.4.2 Proof of Corollary 4

Proof of Corollary 4. Define $\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Gamma}}^\top$ and $\boldsymbol{E}^* = \boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}$. Let $\widehat{\boldsymbol{B}} = \operatorname{argmin}_{\{\boldsymbol{B}\in\mathbb{R}^{p\times d}|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{B}=\boldsymbol{0}\}} \mathcal{L}(\boldsymbol{X}\boldsymbol{B}^\top + \widehat{\boldsymbol{E}})$ and $\widehat{\boldsymbol{\Theta}} = \boldsymbol{X}\widehat{\boldsymbol{B}}^\top + \widehat{\boldsymbol{E}}$. Since $\mathcal{L}(\widehat{\boldsymbol{\Theta}}) \le \mathcal{L}(\boldsymbol{\Theta}^*)$ as assumed in Corollary 4, from Theorem 2 we have $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\| \lesssim \mathcal{O}_{\mathbb{P}}(\sqrt{n \vee p})$. From Theorem 3, we further have

$$\|\widehat{\boldsymbol{E}} - \boldsymbol{E}^*\|_{\mathrm{F}} \le \|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\|_{\mathrm{F}} + \|\boldsymbol{\Theta}^*(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\boldsymbol{\Gamma}^*})\|_{\mathrm{F}} \lesssim \mathcal{O}_{\mathbb{P}}(\sqrt{n \vee p}).$$

Then from Lin et al. [104, Proposition 5.2] we have that, up to sign,

$$\frac{1}{n}\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}^*\boldsymbol{R}^\top\|_{\mathrm{F}}^2 \lesssim \frac{r^{4k_2-k_1+4}}{n \wedge p} =: \eta_n,$$

with probability at least $1 - n^{-c} - p^{-c}$.

Recall the invertible matrix $\boldsymbol{R}$ with $\|\boldsymbol{R}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}(1)$ from Assumption 4 and define the transformed parameters $\widetilde{\boldsymbol{Z}}^* = \boldsymbol{Z}^*\boldsymbol{R}^\top$ and $\widetilde{\boldsymbol{\Gamma}}^* = \boldsymbol{\Gamma}^*\boldsymbol{R}^{-1}$. Because $\boldsymbol{E}_j - \boldsymbol{E}_j^* = \widetilde{\boldsymbol{Z}}^*(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j^*) + (\boldsymbol{Z} - \widetilde{\boldsymbol{Z}}^*)\boldsymbol{\gamma}_j$, with probability tending to one, it follows that

$$\begin{aligned}
\max_{1\le j\le p}\frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{E}}_j - \boldsymbol{E}_j^*\|_2 &= \max_{1\le j\le p}\frac{1}{\sqrt{n}}\|\widetilde{\boldsymbol{Z}}^*(\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j^*) + (\widehat{\boldsymbol{Z}} - \widetilde{\boldsymbol{Z}}^*)\widehat{\boldsymbol{\gamma}}_j\|_2 \\
&\le \max_{1\le j\le p}\frac{1}{\sqrt{n}}\|\widetilde{\boldsymbol{Z}}^*(\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j^*)\|_2 + \max_{1\le j\le p}\frac{1}{\sqrt{n}}\|(\widehat{\boldsymbol{Z}} - \widetilde{\boldsymbol{Z}}^*)\widehat{\boldsymbol{\gamma}}_j\|_2 \\
&\le \max_{1\le j\le p, 1\le i\le n}\frac{1}{\sqrt{n}}\|\widetilde{\boldsymbol{z}}_i^*\|_\infty\|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j^*\|_1 + \max_{1\le j\le p}\frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{Z}} - \widetilde{\boldsymbol{Z}}^*\|_{\mathrm{op}}\|\widehat{\boldsymbol{\gamma}}_j\|_2 \\
&\lesssim \frac{\log n}{\sqrt{n}} + \sqrt{\eta_n} \\
&\lesssim \sqrt{n^{-1}\log n \vee \eta_n}
\end{aligned}$$

where the second last inequality is because that $\|\widetilde{\boldsymbol{\gamma}}_j^*\|_2 \le C$ from Assumption 3, $\|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j^*\|_1 \le \sqrt{r}\|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j^*\|_2 \le 2\sqrt{r}C$ from Lemma A.3.1, and $\widetilde{\boldsymbol{z}}_i^*$'s are independent r-dimensional sub-Gaussian random vectors from Lemma A.4.5 so that $\max_{1\le i\le n}\|\widetilde{\boldsymbol{z}}_i^*\|_\infty$ scales in $\log(nr)$. □

A.4.3 Proof of Theorem 5

Proof of Theorem 5. Define $\boldsymbol{\Delta}_\beta = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\boldsymbol{\Delta}_\zeta = \widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*$, and $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$. Let $\boldsymbol{g}_\beta = \nabla_{\boldsymbol{\beta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$, and $\boldsymbol{g}_\zeta = \nabla_{\boldsymbol{\zeta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$ and analogously define $\boldsymbol{g}_\beta^*, \boldsymbol{g}_\zeta^*$.

(1) Cone condition. From Lemma A.4.1, we have the optimality condition

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + \lambda\|\boldsymbol{\beta}^*\|_1 + \tau_{n,p},$$

with $\tau_{n,p}$ defined in Lemma A.4.1. Rearranging the above display, it follows that

$$\begin{aligned}
\lambda\|\widehat{\boldsymbol{\beta}}\|_1 &\le \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) - \mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda\|\boldsymbol{\beta}^*\|_1 + \tau_{n,p} \\
&\le \boldsymbol{\Delta}_\beta^\top\boldsymbol{g}_\beta^* + \boldsymbol{\Delta}_\zeta^\top\boldsymbol{g}_\zeta^* + \lambda\|\boldsymbol{\beta}^*\|_1 + \tau_{n,p} \\
&\le \|\boldsymbol{\Delta}_\beta\|_1\|\boldsymbol{g}_\beta^*\|_\infty + \boldsymbol{\Delta}_\zeta^\top\boldsymbol{g}_\zeta^* + \lambda\|\boldsymbol{\beta}^*\|_1 + \tau_{n,p}, \qquad \text{(A.4.2)}
\end{aligned}$$

where the second inequality is from the convexity of $\mathcal{L}$, and the last is from Holder's inequality. The term involves nuisance parameters can be further bounded as

$$\boldsymbol{\Delta}_\zeta^\top \boldsymbol{g}_\zeta^* = \frac{1}{n}\sum_{\ell=1}^{np}[\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)]\delta_{\zeta,\ell} \le 2C\|\boldsymbol{g}_\zeta^*\|_\infty \le \frac{4cC^2\alpha\log(2np)}{n}, \tag{A.4.3}$$

where the last inequality is from Lemma A.4.2 (2), which holds with probability at least at least $1-(n+p)^{-c}-(np)^{-c}-\exp(-n)$ for some $c>0$. On the other hand, the left hand side of (A.4.2) is lower bounded as

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}\|_1 &= \|\boldsymbol{\beta}^* + \boldsymbol{\Delta}_\beta\|_1 \\ &= \|\boldsymbol{\beta}^*_{\mathcal{S}} + \boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}_{\beta,\mathcal{S}^c}\|_1 \\ &\ge \|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 - \|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}_{\beta,\mathcal{S}^c}\|_1,\end{aligned} \tag{A.4.4}$$

where $\mathcal{S} = \{j \in [pd] \mid \beta_j^* \neq 0\}$ is the active set of the true coefficients. Combining (A.4.2), (A.4.3), and (A.4.4) yields that

$$(\lambda - \|\boldsymbol{g}_\beta^*\|_\infty)\|\boldsymbol{\Delta}_{\beta,\mathcal{S}^c}\|_1 \le (\lambda + \|\boldsymbol{g}_\beta^*\|_\infty)\|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \left(\frac{4cC^2\alpha\log(2np)}{n} + \tau_{n,p}\right).$$

Note that under the same probabilistic event above, from Lemma A.4.2 (1), we have

$$\|\boldsymbol{g}_\beta^*\|_\infty \le 4\nu^2\sqrt{c\log^2(2nd)/n}. \tag{A.4.5}$$

When $\lambda^* \asymp 8\nu^2\sqrt{c\log^2(2nd)/n} \ge 2\|\boldsymbol{g}_\beta^*\|_\infty$, this implies the approximate cone condition $\boldsymbol{\Delta}_\beta \in \mathcal{C}(3,\mathcal{S})$, where

$$\mathcal{C}(\xi,\mathcal{S}) := \left\{\boldsymbol{\Delta}_\beta \in \mathbb{R}^{pd} \,\middle|\, \|\boldsymbol{\Delta}_{\beta,\mathcal{S}^c}\|_1 \le \xi\|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \tau_{n,p}^*\right\}, \tag{A.4.6}$$

and

$$\tau_{n,p}^* = \frac{C^2\alpha}{\nu^2}\sqrt{\frac{c\log^2(2np)}{n\log^2(2nd)}} + \frac{\sqrt{n}}{(n\wedge p)\log(2nd)} + \sqrt{\frac{(sd)^2}{n\wedge p^{1-k}}}.$$

From the cone condition (A.4.6), the ℓ_1-norm bound follows by observing that

$$\begin{aligned}\|\boldsymbol{\Delta}_\beta\|_1 &\le \|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}_{\beta,\mathcal{S}^c}\|_1 \\ &\le 4\|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_1 + \tau_{n,p}^* \\ &\le 4\sqrt{sd}\|\boldsymbol{\Delta}_{\beta,\mathcal{S}}\|_2 + \tau_{n,p}^* \\ &\le 4\sqrt{sd}\|\boldsymbol{\Delta}_\beta\|_2 + \tau_{n,p}^*.\end{aligned} \tag{A.4.7}$$

(2) Upper and lower bound of the first-order approximation error. To quantify the estimation of the coefficient, we next analyze the first-order approximation error of the normalized

likelihood: $\mathcal{E}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*; \widehat{\boldsymbol{\zeta}}, \boldsymbol{\zeta}^*) = \boldsymbol{\Delta}_\beta^\top (\boldsymbol{g}_\beta - \boldsymbol{g}_\beta^*)$. By the first-order optimality condition of convex optimization problem (2.3.3), we have

$$\begin{aligned}\mathcal{E}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*; \widehat{\boldsymbol{\zeta}}, \boldsymbol{\zeta}^*) &\le -\boldsymbol{\Delta}_\beta^\top \boldsymbol{g}_\beta^* \\ &\le \|\boldsymbol{\Delta}_\beta\|_1 \|\boldsymbol{g}_\beta^*\|_\infty \\ &\le \|\boldsymbol{\Delta}_\beta\|_2 16\nu^2 \sqrt{\frac{csd\log^2(2nd)}{n}} + 4C^2\alpha\sqrt{c}\frac{\log(2np)}{n}, \end{aligned} \tag{A.4.8}$$

where the second inequality is from Holder's inequality, and the last inequality is from (A.4.5) and (A.4.7). This establishes the upper bound for $\mathcal{E}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*; \widehat{\boldsymbol{\zeta}}, \boldsymbol{\zeta}^*)$.

On the other hand, Lemma A.4.3 implies that $\mathcal{E}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*; \widehat{\boldsymbol{\zeta}}, \boldsymbol{\zeta}^*)$ is locally restricted strongly convex over an augmented cone $\mathcal{C}_a(3, \mathcal{S}, \eta_n)$, defined in (A.4.20):

$$\inf_{(\boldsymbol{\Delta}_\beta, \boldsymbol{\Delta}_\zeta) \in \mathcal{C}_a(3,\mathcal{S},\eta_n)} \mathcal{E}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}_\beta, \boldsymbol{\beta}^*; \boldsymbol{\zeta}^* + \boldsymbol{\Delta}_\zeta, \boldsymbol{\zeta}^*) \ge \frac{\kappa_1 C'}{2}\|\boldsymbol{\Delta}_\beta\|_2^2 - C''\sqrt{\frac{\log n}{n} \vee \eta_n}\|\boldsymbol{\Delta}_\beta\|_1, \tag{A.4.9}$$

for some constant $C' > 0$, with probability at least $1 - 2\exp(-3\log n)$.

(3) Estimation error. From Part (2), (A.4.8) and (A.4.9) imply that

$$\begin{aligned}\frac{\kappa_1 C'}{2}\|\boldsymbol{\Delta}_\beta\|_2^2 \le \|\boldsymbol{\Delta}_\beta\|_2 &\left(16\nu^2\sqrt{\frac{csd\log^2(2nd)}{n}} + 4C''\sqrt{\frac{sd}{n\wedge p}}\right) \\ &+ 4C^2\alpha\sqrt{c}\frac{\log(2np)}{n} + \frac{C'' r^{4k_2-k_1+4}}{\sqrt{n\wedge p}}\tau_{n,p}^*.\end{aligned}$$

over $\mathcal{C}_a(3, \mathcal{S}, \eta_n)$. This implies that, with probability at least $1 - (n+p)^{-c'} - (np)^{-c'} - \exp(-n)$ for some $c' > 0$,

$$\begin{aligned}\|\boldsymbol{\Delta}_\beta\|_2 &\le \frac{2}{\kappa_1 C'}\left(16\nu^2\sqrt{\frac{csd\log^2(2nd)}{n}} + C''\sqrt{\frac{sd}{n\wedge p}}\right) \\ &\quad + \sqrt{\frac{2}{\kappa_1 C'}}\sqrt{4C^2\alpha\sqrt{c}\frac{\log(2np)}{n} + \frac{C'' r^{4k_2-k_1+4}\tau_{n,p}^*}{\sqrt{n\wedge p}}} \\ &\lesssim \sqrt{\frac{(sd\log^2(nd)) \vee \log(np)}{n}} + \frac{n^{1/4}}{(n\wedge p)^{3/2}\log^{1/2}(nd)} + \sqrt{\frac{sd}{n\wedge p^{1-k}}}. \end{aligned} \tag{A.4.10}$$

To establish the ℓ_1-norm bound, from (A.4.7) and (A.4.10), we have

$$\begin{aligned}\|\boldsymbol{\Delta}_\beta\|_1 &\le 4\sqrt{sd}\|\boldsymbol{\Delta}_\beta\|_2 + \tau_{n,p}^* \\ &\lesssim \sqrt{sd\frac{(sd\log^2(nd)) \vee \log(np)}{n}} + \sqrt{\frac{(sd)^2}{n\wedge p^{1-k}}} + \frac{\sqrt{n}}{(n\wedge p)\log(nd)}, \end{aligned} \tag{A.4.11}$$

which completes the proof. □

A.4.4 Technical lemmas

Lemma A.4.1 (Sequential and joint optimization)**.** Under the same conditions in Theorem 5, for any constant $\delta > 0$ it holds with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \exp(-n)$ that

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda \|\widehat{\boldsymbol{\beta}}\|_1 \leq \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + \lambda \|\boldsymbol{\beta}^*\|_1 + \tau_{n,p},$$

where

$$\tau_{n,p} = \mathcal{O}\left(\frac{1}{n \wedge p} + \lambda \sqrt{\frac{(sd)^2}{n \wedge p^{1-k}}} \right).$$

Proof of Lemma A.4.1. From Assumption 1, the entry of $\boldsymbol{B}^*$ is bounded because $\|\boldsymbol{B}^*\|_{\max} = \max_{1 \leq i \leq p, 1 \leq j \leq d} |b_{ij}| \leq \max_{1 \leq i \leq p} \|\boldsymbol{b}_i\|_\infty \leq \max_{1 \leq i \leq p} \|\boldsymbol{b}_i\|_2 \leq C$. From Proposition 1 and Assumption 1, we further have that

$$\|\mathcal{P}_{\boldsymbol{\Gamma}^*} \boldsymbol{B}^*\|_{\mathrm{F}} \lesssim \sqrt{sd/p} \|\boldsymbol{B}^*\|_{\max} \lesssim \sqrt{sd/p}. \tag{A.4.12}$$

Thus, from the assumption of Theorem 5, we have that

$$\|\mathcal{P}_{\boldsymbol{\Gamma}^*} \boldsymbol{B}^*\|_{1,1} = \mathcal{O}(p^{k/2} \|\mathcal{P}_{\boldsymbol{\Gamma}^*} \boldsymbol{B}^*\|_{\mathrm{F}}) \lesssim \sqrt{sd} p^{(k-1)/2}. \tag{A.4.13}$$

This ensures that $\|\widetilde{\boldsymbol{\beta}}^*\|_{1,1}$ is close to $\|\boldsymbol{\beta}^*\|_{1,1}$.

From optimality condition of optimization problem (2.3.5), we have

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda \|\widehat{\boldsymbol{\beta}}\|_1 \leq \mathcal{L}(\widetilde{\boldsymbol{\beta}}', \widetilde{\boldsymbol{\zeta}}') + \lambda \|\widetilde{\boldsymbol{\beta}}'\|_1,$$

where $\widetilde{\boldsymbol{\beta}}' = \mathrm{vec}(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp} \boldsymbol{B}^*)$ and $\widetilde{\boldsymbol{\zeta}}' = \mathrm{vec}(\boldsymbol{\Theta}^* \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}})$. It follows that

$$\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda \|\widehat{\boldsymbol{\beta}}\|_1 \leq \mathcal{L}(\boldsymbol{\Theta}^* - \boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}) + \lambda \|\boldsymbol{B}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{1,1}. \tag{A.4.14}$$

We next bound the first term in (A.4.14). From the proof of Lemma A.2.1, we have

$$\mathcal{L}(\boldsymbol{\Theta}^* - \boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}) - \mathcal{L}(\boldsymbol{\Theta}^*) \leq \underbrace{\frac{1}{n} \mathrm{tr}((\boldsymbol{Y} - A'(\boldsymbol{\Theta}^*))^\top \boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp})}_{T_1} + \underbrace{\frac{\kappa_2}{2n} \|\boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{\mathrm{F}}^2}_{T_2}.$$

Let $\boldsymbol{A} = n^{-1} \boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}$. From Lemmas A.4.4 and A.4.5, the second term T_2 can be upper bounded as

$$T_2 = \frac{\kappa_2}{2} n \|\boldsymbol{A}\|_{\mathrm{F}}^2 = \frac{\kappa_2}{2n} \|\boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{\mathrm{F}}^2 \leq \frac{\kappa_2}{2n} \|\boldsymbol{Z}^*\|_{\mathrm{op}}^2 \|\boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{\mathrm{F}}^2 = \mathcal{O}\left(\frac{r}{n \wedge p} \right),$$

with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \exp(-n)$. In the equality, we use the fact that $\boldsymbol{Z}^*$ is a matrix with independent sub-Gaussian rows to obtain similar results as Lemma A.3.2.

For the first term T_1, note that $y_{ij} - A'(\theta_{ij}^*)$ is mean-zero (ν, α)-sub-exponential random variable when conditioned on $(\boldsymbol{X}^*, \boldsymbol{Z}^*)$, where $\nu = \sqrt{\kappa_2}$ and $\alpha = 1/C^2$, as shown in the proof of Theorem 2. To bound the first term T_1, we apply Bernstein's inequality [172, Theorem 2.8.2] to obtain

$$\mathbb{P}(|T_1| \geq t \mid \boldsymbol{X}^*, \boldsymbol{Z}^*) \leq 2 \exp\left(- \min\left\{ \frac{t^2}{2\nu^2 \|\boldsymbol{A}\|_{\mathrm{F}}^2}, \frac{t}{2\alpha \|\boldsymbol{A}\|_{\max}} \right\} \right)$$

From the proof above, we have that $\|\boldsymbol{A}\|_{\max} \leq \|\boldsymbol{A}\|_{\mathrm{F}} = \mathcal{O}_{\mathbb{P}}((nr(n \wedge p))^{-1/2})$. Therefore, by choosing $t \asymp (nr(n \wedge p))^{-1/2} \log(np)$, we further have

$$|T_1| = \mathcal{O}\left(\sqrt{\frac{r}{n(n \wedge p)}}\right).$$

with probability at least $1 - (np)^{-\delta}$. The above results imply that

$$\mathcal{L}(\boldsymbol{\Theta}^* - \boldsymbol{Z}^* \boldsymbol{\Gamma}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}) \leq \mathcal{L}(\boldsymbol{\Theta}^*) + \mathcal{O}\left(\frac{r}{n \wedge p}\right), \tag{A.4.15}$$

with probability at least $1 - (n+p)^{-\delta} - (np)^{-\delta} - \exp(-n)$.

Consider the second term of (A.4.14), from Theorem 3 we also have

$$\begin{aligned}
\|\boldsymbol{B}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{1,1} - \|\boldsymbol{B}^{*\top} \mathcal{P}_{\boldsymbol{\Gamma}}^{*\perp}\|_{1,1} &\leq \|\boldsymbol{B}^{*\top}(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp} - \mathcal{P}_{\boldsymbol{\Gamma}}^{*\perp})\|_{1,1} \\
&\leq \sum_{\ell=1}^{d} \|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp} - \mathcal{P}_{\boldsymbol{\Gamma}}^{*\perp}) \boldsymbol{B}_{\ell}^*\|_1 \\
&\leq dsC \max_{j \in [d]} \|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp} - \mathcal{P}_{\boldsymbol{\Gamma}}^{*\perp}) \boldsymbol{e}_j\|_1 \\
&= \mathcal{O}\left(\frac{ds}{\sqrt{p(n \wedge p)}}\right),
\end{aligned} \tag{A.4.16}$$

under the same probability event above.

Finally, combining (A.4.14)-(A.4.16), we have

$$\begin{aligned}
\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}) + \lambda \|\widehat{\boldsymbol{\beta}}\|_1 &\leq \mathcal{L}(\boldsymbol{\Theta}^*) + \lambda \|\boldsymbol{B}^{*\top} \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\|_{1,1} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n \wedge p}\right), \\
&= \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*, \widetilde{\boldsymbol{\zeta}}^*) + \lambda \|\widetilde{\boldsymbol{\beta}}^*\|_1 + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n \wedge p}\right) \\
&= \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) + \lambda \|\boldsymbol{\beta}^*\|_1 + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n \wedge p} + \lambda \sqrt{\frac{(sd)^2}{n \wedge p^{1-k}}}\right),
\end{aligned}$$

where the last inequality is from (A.4.13). This completes the proof. □

Lemma A.4.2 (Infinity norm of the gradient)**.** Under the same conditions in Theorem 5, for any constant $c > 0$ it holds that

(1) $\|\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \leq 4\nu^2 \sqrt{\frac{c \log^2(2nd)}{n}}$,

(2) $\|\nabla_{\boldsymbol{\zeta}} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \leq \frac{2c\alpha \log(2np)}{n}$,

with probability at least $1 - (2nd)^{-c} - (2np)^{-c}$.

Proof of Lemma A.4.2. Recall that

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) = \frac{1}{n} \sum_{\ell=1}^{np} [\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)] \widetilde{\boldsymbol{x}}_\ell \tag{A.4.17}$$

$$\nabla_{\zeta_\ell} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) = \frac{1}{n} [\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)], \qquad \ell \in [np], \tag{A.4.18}$$

where $\widetilde{\boldsymbol{x}}_\ell$ is the ℓ-th row of $\widetilde{\boldsymbol{X}}$. We split the proof into different parts.

Part (1). Conditioned on $\boldsymbol{X}$ and $\boldsymbol{Z}^*$, the term $\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)$ is a zero-mean (ν, α)-sub-exponential random variable, where $\nu = \sqrt{\kappa_2}$ and $\alpha = 1/C^2$, as shown in the proof of Theorem 2. Let $C' = \max_{\ell \in [np]} \|\widetilde{\boldsymbol{x}}_\ell^*\|_\infty$. Because $\widetilde{\boldsymbol{x}}_\ell^*$ are sparse vectors with $\|\widetilde{\boldsymbol{x}}_\ell^*\|_0 = d$, we have that $[\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)]\widetilde{x}_{\ell j}$ is a zero-mean $(\nu, C'\alpha)$-sub-exponential random variable for $j = k, k+p, \ldots, k+p(d-1)$ and zero otherwise, where $k = \lfloor p/d \rfloor$. By Bernstein's inequality, it follows that

$$\mathbb{P}\left(\frac{1}{n}\sum_{\ell=1}^{np}[\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)]\widetilde{x}_{\ell j} \le t\right) \ge 1 - 2\exp\left(-\frac{n}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{C'\alpha}\right\}\right).$$

Applying union bound over $j = k, k+p, \ldots, k+p(d-1)$ yields that

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le t\right) \ge 1 - 2d\exp\left(-n\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2C'\alpha}\right\}\right).$$

By setting $t = C'\sqrt{2\nu^2 c\log(2nd)/n}$ for some fixed constant $c > 1$, we have

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le C'\sqrt{\frac{2\nu^2 c\log(2nd)}{n}}\right) \ge 1 - (2nd)^{1-c},$$

when n is large enough such that $t < \nu^2/(C'\alpha)$. By Lemma A.4.6, we also have $C' = \|\widetilde{\boldsymbol{X}}\|_{\max} \le 2\sqrt{2}\nu\sqrt{c\log(nd)}$ with probability at least $1 - (2nd)^{1-c}$. It follows that

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le 4\nu^2 c\sqrt{\frac{\log^2(2nd)}{n}}\right) \ge 1 - 2(2nd)^{1-c}, \tag{A.4.19}$$

which finishes the proof for Part (1).

Part (2). Because $\widetilde{y}_\ell - A'(\widetilde{\boldsymbol{x}}_\ell^\top \boldsymbol{\beta}^* + \zeta_\ell^*)$'s are zero-mean (ν, α)-sub-exponential random variables, by union bound, we have

$$\mathbb{P}\left(n\|\nabla_{\boldsymbol{\zeta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le t\right) \ge 1 - 2np\exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right),$$

or equivalently,

$$\begin{aligned}\mathbb{P}\left(\|\nabla_{\boldsymbol{\zeta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le t\right) &\ge 1 - 2np\exp\left(-\frac{1}{2}\min\left\{\frac{(nt)^2}{\nu^2}, \frac{nt}{\alpha}\right\}\right) \\ &= 1 - 2np\exp\left(-\frac{nt}{2\alpha}\right)\end{aligned}$$

when $t \ge \nu^2/(\alpha n)$ is sufficiently large. Choosing $t = \max\{2c\alpha\log(2np), \nu^2/\alpha\}/n$ for any constant $c > 1$ yields that

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\zeta}}\mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)\|_\infty \le \frac{2c\alpha\log(2np)}{n}\right) \le 1 - (2np)^{1-c}.$$

Finally, we obtain the tail probability as the lemma states by taking the union bound over the above events. □

Lemma A.4.3 (Locally restricted strongly convexity)**.** Under the same conditions in Theorem 5, define the augmented cone

$$\begin{aligned}\mathcal{C}_a(\xi, \mathcal{S}, \eta_n) := \{(\boldsymbol{\Delta}_\beta, \boldsymbol{\Delta}_\zeta) \in \mathbb{R}^{pd} \times \mathbb{R}^{np} \mid \boldsymbol{\Delta}_\beta \in \mathcal{C}(\xi, \mathcal{S}), \\ \max_{j\in[p]} \frac{1}{n}\|\boldsymbol{Z}\boldsymbol{\gamma}_j - \boldsymbol{Z}^*\boldsymbol{\gamma}_j^*\|_2^2 \le \frac{\log n}{n} \vee \eta_n, \text{ such that } \boldsymbol{\zeta} = \mathrm{vec}(\boldsymbol{Z}\boldsymbol{\Gamma}^\top)\},\end{aligned} \tag{A.4.20}$$

where the cone $\mathcal{C}(\xi, \mathcal{S})$ defined in (A.4.6) and $\eta_n = o(1)$. Then, it holds that

$$\inf_{(\boldsymbol{\Delta}_\beta, \boldsymbol{\Delta}_\zeta)\in\mathcal{C}_a(\xi,\mathcal{S},\eta_n)} \mathcal{E}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}_\beta, \boldsymbol{\beta}^*; \boldsymbol{\zeta}^* + \boldsymbol{\Delta}_\zeta, \boldsymbol{\zeta}^*) \ge \frac{\kappa_1 C'}{2}\|\boldsymbol{\Delta}_\beta\|_2^2 - C''\sqrt{n^{-1} \vee \eta_n}\|\boldsymbol{\Delta}_\beta\|_1,$$

for some constant $C' > 0$, with probability at least $1 - 2\exp(-\xi \log n)$.

Proof of Lemma A.4.3. For all $(\boldsymbol{\Delta}_\beta, \boldsymbol{\Delta}_\zeta) \in \mathcal{C}_a(\xi, \mathcal{S}, \eta_n)$, let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{\Delta}_\beta$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}^* + \boldsymbol{\Delta}_\zeta$. Let $\boldsymbol{\Delta}_{\beta_j} \in \mathbb{R}^d$ be the sub-vector containing the $(j-1)d$-th to $(jd-1)$-th entries of $\boldsymbol{\Delta} \in \mathbb{R}^{pd}$. It is also equivalent to $\boldsymbol{b}_j - \boldsymbol{b}_j^*$, the j-th row of $\mathrm{vec}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \in \mathbb{R}^{p\times d}$. Let $\boldsymbol{E} = \mathrm{vec}^{-1}(\boldsymbol{\zeta}) \in \mathbb{R}^{n\times p}$ and $\boldsymbol{E}^* = \mathrm{vec}^{-1}(\boldsymbol{\zeta}^*) \in \mathbb{R}^{n\times p}$.

We begin by decomposing the error into different terms:

$$\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^*; \boldsymbol{\zeta}, \boldsymbol{\zeta}^*) = \sum_{j=1}^p \mathcal{E}_j(\boldsymbol{\beta}, \boldsymbol{\beta}^*; \boldsymbol{\zeta}, \boldsymbol{\zeta}^*),$$

where

$$\begin{aligned}\mathcal{E}_j(\boldsymbol{\beta}, \boldsymbol{\beta}^*; \boldsymbol{\zeta}, \boldsymbol{\zeta}^*) &= \frac{1}{n}\sum_{i=1}^n [A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j) - A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^{*\top} \boldsymbol{\gamma}_j^*)]\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j} \\ &= \frac{1}{n}\sum_{i=1}^n [A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j) - A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j)]\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j} \\ &\quad + \frac{1}{n}\sum_{i=1}^n [A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j) - A'(\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^{*\top} \boldsymbol{\gamma}_j^*)]\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j} \\ &= \frac{1}{n}\sum_{i=1}^n A''(\theta_{ij})(\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j})^2 + \frac{1}{n}\sum_{i=1}^n A''(\theta'_{ij})(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j - \boldsymbol{z}_i^{*\top} \boldsymbol{\gamma}_j^*)\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j} \\ &= T_{1j} + T_{2j},\end{aligned} \tag{A.4.21}$$

where θ_{ij} is between $\boldsymbol{x}_i^\top \boldsymbol{b}_j + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j$ and $\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j$, and θ'_{ij} is between $\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j$ and $\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + \boldsymbol{z}_i^{*\top} \boldsymbol{\gamma}_j^*$.

For the first term, because by Assumption 1, $\kappa_1 := \inf_{\theta\in\mathcal{R}} A''(\theta) > 0$, we have

$$\begin{aligned}T_{1j} &\ge \kappa_1 \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{\Delta}_{\beta_j})^2 \\ &\ge \kappa_1 \lambda_p\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)\|\boldsymbol{\Delta}_{\beta_j}\|_2^2 \\ &\ge \frac{\kappa_1}{C}\lambda_p\left(\frac{1}{n}\boldsymbol{\Sigma}_x^{-\frac{1}{2}}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\Sigma}_x^{-\frac{1}{2}}\right)\|\boldsymbol{\Delta}_{\beta_j}\|_2^2,\end{aligned}$$

where the last inequality is from Assumption 2. From Vershynin [172, Lemma 10.6.6], we further have that when $n \gtrsim \xi \log d$

$$T_{1j} \geq \frac{\kappa_1 C'}{2} \|\boldsymbol{\Delta}_{\beta_j}\|_2^2 \tag{A.4.22}$$

over $\mathcal{C}(\xi, \mathcal{S})$ (which contains $\mathcal{C}_a(\xi, \mathcal{S}, C_0)$), with probability at least $1 - 2\exp(-\xi \log n)$ for some absolute constant $C' > 0$.

For the second term, by Holder's inequality, we have

$$\begin{aligned} |T_{2j}| &= \left| \frac{1}{n} (\boldsymbol{E}_j - \boldsymbol{E}_j^*)^\top \mathrm{diag}(A''(\boldsymbol{\Theta}_j')) \boldsymbol{X} \boldsymbol{\Delta}_{\beta_j} \right| \\ &\leq \frac{1}{n} \|\boldsymbol{X}^\top \mathrm{diag}(A''(\boldsymbol{\Theta}_j'))(\boldsymbol{E}_j - \boldsymbol{E}_j^*)\|_\infty \cdot \|\boldsymbol{\Delta}_{\beta_j}\|_1 \\ &= \max_{1 \leq \ell \leq d} \frac{1}{n} |(\boldsymbol{E}_j - \boldsymbol{E}_j^*)^\top \mathrm{diag}(A''(\boldsymbol{\Theta}_j')) \boldsymbol{X}_\ell| \cdot \|\boldsymbol{\Delta}_{\beta_j}\|_1, \end{aligned} \tag{A.4.23}$$

where $\boldsymbol{\Theta}_j' = (\theta_{1j}', \ldots, \theta_{nj}')^\top$. Recall that $\boldsymbol{E} = \boldsymbol{Z}\boldsymbol{\Gamma}^\top$ and $\boldsymbol{E}^* = \boldsymbol{Z}^*\boldsymbol{\Gamma}^{*\top}$ are such that $\max_{1 \leq j \leq d} \frac{1}{\sqrt{n}} \|\boldsymbol{E}_j - \boldsymbol{E}_j^*\|_2 \lesssim \sqrt{(n^{-1} \log n) \vee \eta_n}$. From Corollary 4, we have

$$\begin{aligned} |T_{2j}| &\leq \kappa_2 \max_{1 \leq j \leq d} \frac{1}{\sqrt{n}} \|\boldsymbol{E}_j - \boldsymbol{E}_j^*\|_2 \cdot \frac{1}{\sqrt{n}} \|\boldsymbol{X}\|_{\mathrm{op}} \cdot \|\boldsymbol{\Delta}_{\beta_j}\|_1 \\ &\lesssim \sqrt{(n^{-1} \log n) \vee \eta_n} \|\boldsymbol{\Delta}_{\beta_j}\|_1, \end{aligned} \tag{A.4.24}$$

where the last inequality is because $\|\boldsymbol{X}\|_{\mathrm{op}} \lesssim \sqrt{n}$.

By combining (A.4.21), (A.4.22) and (A.4.24), we have that

$$\mathcal{E}_j(\boldsymbol{\beta}, \boldsymbol{\beta}^*; \boldsymbol{\zeta}, \boldsymbol{\zeta}^*) \geq \frac{\kappa_1 C'}{2} \|\boldsymbol{\Delta}_{\beta_j}\|_2^2 - C'' \sqrt{n^{-1} \vee \eta_n} \|\boldsymbol{\Delta}_{\beta_j}\|_1,$$

for some constant $C'' > 0$. Thus,

$$\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^*; \boldsymbol{\zeta}, \boldsymbol{\zeta}^*) \geq \frac{\kappa_1 C'}{2} \|\boldsymbol{\Delta}_{\beta}\|_2^2 - C'' \sqrt{n^{-1} \vee \eta_n} \|\boldsymbol{\Delta}_{\beta}\|_1,$$

over the cone $\mathcal{C}_a(\xi, \mathcal{S}, \eta_n)$. □

Remark 12 (Neyman orthogonality)**.** By coincidence, the proof above on T_{2j} actually verifies the uniform Neyman orthogonality of the empirical loss in semiparametric models [33, 53]. This relies on the estimation error rates for the nuisance parameters $\boldsymbol{E}_j^*$ by using the consistent estimation for the latent factors $\boldsymbol{Z}^*$ with rate η_n as assumed in the cone condition (A.4.21). To see this, recall that the pathwise derivative map of the gradient $\nabla_{\boldsymbol{b}_j} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$ evaluated at the true parameter $\boldsymbol{\beta}^*$ and nuisance component value $\boldsymbol{\zeta}^*$ (when $t = 0$) is given by

$$\begin{aligned} \frac{\partial}{\partial t} \nabla_{\boldsymbol{b}_j} \mathcal{L}(\boldsymbol{\beta}^*, t(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*) + \boldsymbol{\zeta}^*)\Big|_{t=0} &= \frac{1}{n} \sum_{i \in [n]} A''(\boldsymbol{x}_i^\top \boldsymbol{b}_j^* + e_{ij}^*)(e_{ij} - e_{ij}^*) \boldsymbol{x}_i \\ &= \frac{1}{n} \boldsymbol{X}^\top \mathrm{diag}(A''(\boldsymbol{\Theta}_j^*))(\boldsymbol{E}_j - \boldsymbol{E}_j^*). \end{aligned}$$

Compared to (A.4.23), up to a constant factor, (A.4.24) also suggest that the pathwise derivative's infinity norm vanishes with a rate of $\sqrt{n^{-1} \vee \eta_n}$. In other words, at the true parameter value, local perturbations of the nuisance component around its true value have a negligible effect on the gradient of the loss with respect to the primary parameter, with high probability; see [33, 53] for more detailed discussions about the Neyman orthogonality.

Lemma A.4.4 (Bounds related to projection)**.** Under assumptions in Theorem 5, it holds that

$$\|\mathcal{P}_{\widehat{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}^*\|_{\mathrm{F}}^2 = \mathcal{O}\left(\frac{r}{n \wedge p}\right),$$

with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$.

Proof of Lemma A.4.4. From the proof of Theorem 3, the result follows by noting that

$$\begin{aligned}
\|\mathcal{P}_{\widehat{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}^*\|_{\mathrm{F}}^2 &= \|(\mathcal{P}_{\widehat{\mathbf{\Gamma}}}^{\perp} - \mathcal{P}_{\mathbf{\Gamma}^*}^{\perp})\mathbf{\Gamma}^*\|_{\mathrm{F}}^2 \\
&= \sum_{\ell=1}^{r}\left\|\sum_{j=1}^{p}(\mathcal{P}_{\widehat{\mathbf{\Gamma}}}^{\perp} - \mathcal{P}_{\mathbf{\Gamma}^*}^{\perp})\boldsymbol{e}_j \cdot \gamma_{j,\ell}^*\right\|_2^2 \\
&\leq rpC^2 \max_{1\leq \ell \leq p}\|(\mathcal{P}_{\widehat{\mathbf{\Gamma}}}^{\perp} - \mathcal{P}_{\mathbf{\Gamma}^*}^{\perp})\boldsymbol{e}_j\|_2^2 \\
&= \mathcal{O}(r(n\wedge p)^{-1}),
\end{aligned}$$

with probability at least $1-(n+p)^{-\delta}-(np)^{-\delta}-\exp(-n)$. □

Lemma A.4.5 (Sub-Gaussianity of $\boldsymbol{Z}$)**.** Under Assumptions 1–3, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are independent and identically distributed sub-Gaussian random vectors.

Proof of Lemma A.4.5. Recall that $\boldsymbol{z}_i^* = \boldsymbol{D}^*\boldsymbol{x}_i^* + \boldsymbol{w}_i^*$ is the linear function of two independent sub-Gaussian random vectors. The mean is given by $\mathbb{E}[\boldsymbol{z}_i^*] = \boldsymbol{D}^*\mathbb{E}[\boldsymbol{x}_i^*]$. It suffices to bound the operator norm of $\boldsymbol{D}^*$. From $\mathbf{\Theta}^* = \boldsymbol{X}\boldsymbol{B}^{*\top} + \boldsymbol{Z}^*\mathbf{\Gamma}^{*\top}$, we have $\boldsymbol{X}^\top\mathbf{\Theta}^* = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}^{*\top} + \boldsymbol{X}^\top\boldsymbol{Z}^*\mathbf{\Gamma}^{*\top}$. Taking expectation over $\boldsymbol{X}$ and $\boldsymbol{Z}^*$ yield that $\mathbb{E}[\boldsymbol{X}^\top\mathbf{\Theta}^*/n] = \mathbf{\Sigma}_x(\boldsymbol{B}^{*\top} + \boldsymbol{D}^{*\top}\mathbf{\Gamma}^{*\top})$. Rearranging the formula yields that

$$\mathbf{\Gamma}^*\boldsymbol{D}^* = \frac{1}{n}\mathbb{E}[\mathbf{\Theta}^{*\top}\boldsymbol{X}]\mathbf{\Sigma}_x^{-1} - \boldsymbol{B}^*$$

and

$$\boldsymbol{D}^* = (\mathbf{\Gamma}^{*\top}\mathbf{\Gamma}^*)^{-1}\mathbf{\Gamma}^{*\top}\left(\frac{1}{n}\mathbb{E}[\mathbf{\Theta}^{*\top}\boldsymbol{X}^*]\mathbf{\Sigma}_x^{-1} - \boldsymbol{B}^*\right).$$

By the sub-multiplicative property of the operator norm, we have

$$\begin{aligned}
\|\boldsymbol{D}^*\|_{\mathrm{op}} &\leq \|(\mathbf{\Gamma}^{*\top}\mathbf{\Gamma}^*)^{-1}\mathbf{\Gamma}^{*\top}\|_{\mathrm{op}}\left(\frac{1}{n}\|\mathbb{E}[\mathbf{\Theta}^{*\top}\boldsymbol{X}]\|_{\mathrm{op}}\|\mathbf{\Sigma}_x^{-1}\|_{\mathrm{op}} + \|\boldsymbol{B}^*\|_{\mathrm{op}}\right) \\
&\lesssim \frac{1}{\sqrt{p}}\left(\frac{1}{n}\mathbb{E}[\|\mathbf{\Theta}^{*\top}\|_{\mathrm{op}}\|\boldsymbol{X}\|_{\mathrm{op}}] + \sqrt{p}\right) \\
&\lesssim \frac{1}{\sqrt{p}}\left(\frac{\sqrt{np}\sqrt{n}}{n} + \sqrt{p}\right) \\
&\lesssim 1,
\end{aligned}$$

where in the second inequality, we use Jensen's inequality and the norm inequality $\|\boldsymbol{B}\|_{\mathrm{op}} \leq \sqrt{p}\|\boldsymbol{B}\|_{\max}$. This finishes the proof. □

Lemma A.4.6 (Infinity norm of the covariates)**.** Under Assumption 2, it holds that $\|\widetilde{\boldsymbol{X}}\|_{\max} \leq 2\sqrt{2}\nu\sqrt{c\log(2nd)}$, with probability at least $1-(2nd)^{-c}$ for any fixed constant $c > 0$.

Proof of Lemma A.4.6. Let $\mathrm{vec}^{-1}(\boldsymbol{e}_\ell) = \boldsymbol{e}_{n,i}\boldsymbol{e}_{p,j}^\top$ where $\boldsymbol{e}_{n,i}$ is the unit vector of dimension n and $\boldsymbol{e}_{p,j}$ is defined analogously. Note that

$$\begin{aligned}\widetilde{\boldsymbol{x}}_\ell &= (\boldsymbol{I}_p \otimes \boldsymbol{X})^\top \boldsymbol{e}_\ell \\ &= \mathrm{vec}(\boldsymbol{X}^\top \mathrm{vec}^{-1}(\boldsymbol{e}_\ell)\boldsymbol{I}_p) \\ &= \mathrm{vec}(\boldsymbol{X}^\top \boldsymbol{e}_{n,i}\boldsymbol{e}_{p,j}^\top) \\ &= \mathrm{vec}(\boldsymbol{x}_i \boldsymbol{e}_{p,j}^\top).\end{aligned}$$

Because $\boldsymbol{x}_i$'s are ν-sub-Gaussian random vectors, we have

$$\begin{aligned}\|\widetilde{\boldsymbol{X}}\|_{\max} &= \max_{\ell\in[np]} \|\widetilde{\boldsymbol{x}}_\ell\|_\infty \\ &= \max_{i\in[n],j\in[p]} \|\boldsymbol{x}_i \boldsymbol{e}_{p,j}^\top\|_\infty \\ &\le \max_{i\in[n]} \|\boldsymbol{x}_i\|_\infty \\ &\le \max_{i\in[n],j\in[p]} |x_{ij}| \\ &\le 2\sqrt{2}\nu\sqrt{c\log(2nd)}\end{aligned}$$

with probability at least $1-(2nd)^{-c}$ for any fixed constant $c>0$. □

A.5 Asymptotic normality of the debiased estimator

A.5.1 Proof of Theorem 6

Proof of Theorem 6. Condition on $\mathcal{D}_2$, Theorem 3 and Theorem 5 imply that the following event holds:

$$\begin{aligned}\mathcal{E}_1 = \Big\{ &\max_{1\le j\le p} \|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathcal{P}_{\boldsymbol{\Gamma}^*})\boldsymbol{e}_j\|_2 \lesssim (p(n\wedge p))^{-1/2}, \frac{1}{n}\|\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{Z}^*\boldsymbol{\gamma}_j^*\|_2^2 \lesssim r_{n,p}, \\ &\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{1,1} \lesssim \sqrt{sd}r_{n,p}, \|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{\mathrm{F}} \lesssim r_{n,p}\Big\}.\end{aligned}$$

where $r_{n,p}$ is defined in Theorem 5.

Recall that

$$\widehat{b}_{j1}^{\mathrm{de}} = \widehat{b}_{j1} + \widehat{\boldsymbol{u}}^\top \frac{1}{n}\sum_{i=1}^n \widehat{\omega}_i \boldsymbol{x}_i (\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i))^\top \boldsymbol{v}_i \tag{A.5.1}$$

where $\boldsymbol{v}_i = \mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp \boldsymbol{e}_j$. By Taylor expansion of $A'(\theta_{ij}^*)$ at $\widehat{\theta}_{ij} := \boldsymbol{x}_i^\top \widehat{\boldsymbol{b}}_j + \widehat{\boldsymbol{z}}_i^\top \widehat{\boldsymbol{\gamma}}_j$, we have

$$A'(\theta_{ij}^*) = A'(\widehat{\theta}_{ij}) + A''(\widehat{\theta}_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij}) + \frac{1}{2}A'''(\psi_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij})^2,$$

for some ψ_{ij} between $\widehat{\theta}_{ij}$ and θ_{ij}^*. Then, the residual of the ith sample can be decomposed into three sources of errors:

$$\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i) = \underbrace{\boldsymbol{\epsilon}_i}_{\text{stochastic error}} + \underbrace{\boldsymbol{p}_i}_{\text{remaining bias}} + \underbrace{\boldsymbol{q}_i}_{\text{approximation error}} \tag{A.5.2}$$

where the three terms of errors read that

$$\begin{aligned}\boldsymbol{\epsilon}_i &= \boldsymbol{y}_i - A'(\boldsymbol{\theta}_i^*)\\ \boldsymbol{p}_i &= A''(\widehat{\boldsymbol{\theta}}_i) \odot (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*)\\ \boldsymbol{q}_i &= -\frac{1}{2}[A'''(\psi_{ij})(\theta_{ij}^* - \widehat{\theta}_{ij})^2]_{j\in[p]}.\end{aligned}$$

Substituting (A.5.2) into (A.5.1) yields that

$$\begin{aligned}\widehat{b}_{j1}^{\mathrm{de}} - b_{j1}^* &= (\widehat{b}_{j1} - b_{j1}^*) + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i \boldsymbol{x}_i \boldsymbol{\epsilon}_i^\top \boldsymbol{v}_i + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i \boldsymbol{x}_i \boldsymbol{p}_i^\top \boldsymbol{v}_i + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i \boldsymbol{x}_i \boldsymbol{q}_i^\top \boldsymbol{v}_i\\ &= \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i \boldsymbol{x}_i \boldsymbol{\epsilon}_i^\top \boldsymbol{v}_i + \left(\boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{e}_1^\top\right)(\boldsymbol{b}_j^* - \widehat{\boldsymbol{b}}_j)\\ &\quad + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{B}^{*\top}((\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp - \mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp)\boldsymbol{e}_j - \mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{e}_j)\\ &\quad + \boldsymbol{u}^\top \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{q}_i^\top \boldsymbol{v}_i, \end{aligned} \tag{A.5.3}$$

In the second equality above, we use the properties that $\widehat{\boldsymbol{B}}$ and $\boldsymbol{B}^*$ are in the column spaces of the orthogonal projections $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp$ and $\mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp$, respectively. Denote the four terms in the right-hand side of (A.5.3) by $T_{1j}, T_{2j}, T_{3j}, T_{4j}$, respectively. We will analyze each of them separately, conditioning on $\mathcal{D}_2$ and $\boldsymbol{X}$. Then the randomness is from $\boldsymbol{\epsilon}_i$'s and $\boldsymbol{u}$. We will show that the T_1 inherits $\sqrt{n}$-convergence rate and is asymptotically normally distributed, while the others have faster convergence rates.

Part (1) T_{1j}. From Lemma A.5.2, it follows that

$$\sqrt{n}\frac{T_{1j}}{\sigma_j} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Part (2) T_{2j}. By Holder's inequality and the constraint of optimization problem (2.4.4), it follows that

$$|T_{2j}| \le \lambda_n \|\boldsymbol{b}_j^* - \widehat{\boldsymbol{b}}_j\|_1 \lesssim \sqrt{\frac{\log(nd)}{n}} r'_{n,p},$$

with probability tending to one. Thus, we have $\sqrt{n}|T_{2j}| \xrightarrow{\mathrm{p}} 0$ as $n, p \to \infty$.

Part (3) T_{3j}. From Theorem 3 and (A.1.1) in the proof of Proposition 1, we have that

$$\begin{aligned}|T_{3j}| &\le \|\boldsymbol{u}\|_2 \cdot \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top\right\|_{\mathrm{op}} (\|\boldsymbol{B}^*\|_{\mathrm{op}} \|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp - \mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp)\boldsymbol{e}_j\|_2 + \|\boldsymbol{B}^{*\top}\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{e}_j\|_2)\\ &\lesssim \sqrt{\frac{d}{p(n\wedge p)}} + \frac{sd}{p},\end{aligned}$$

with probability tending to one. Thus, if $\sqrt{n}/p \to 0$, we have $\sqrt{n}|T_{3j}| \xrightarrow{\mathrm{p}} 0$ as $n, p \to \infty$.

Part (4) T_{4j}. The higher-order term is bounded as below:

$$\begin{aligned}|T_{4j}| &\leq \frac{\kappa_2}{2\kappa_1}\max_{1\leq i\leq n}|\boldsymbol{u}^\top\boldsymbol{x}_i|\cdot\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n|\boldsymbol{x}_i^\top(\widehat{\boldsymbol{b}}_j-\boldsymbol{b}_j^*)|^2\cdot\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j\|_2\\ &\lesssim \max_{1\leq i\leq n}|\boldsymbol{u}^\top\boldsymbol{x}_i|^3\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_{\mathrm{F}}^2\\ &\lesssim \tau_n^3 r_{n,p}^2,\end{aligned}$$

with probability tending to one. Thus, if $n/(\log(nd)p^{3/2})\to 0$ and $\sqrt{n}/p^{1-k}\to 0$, we have $\sqrt{n}|T_{4j}|\xrightarrow{\mathrm{p}}0$ as $n,p\to\infty$.

We are now combining the above four terms. Because when $n/\log(nd)=o(p^{3/2})$ and $n=o(p^{2(1-k)})$, $T_{2j},\ldots,T_{4j}=o_{\mathbb{P}}(1/\sqrt{n})$, we have $\sqrt{n}(\widehat{b}_{j1}^{\mathrm{de}}-b_{j1}^*)/\sigma_j\xrightarrow{\mathrm{d}}\mathcal{N}(0,1)$. □

A.5.2 Proof of Proposition 7

Proof of Proposition 7. By the definition of t_j and (A.5.3) we have the decomposition

$$t_j=\vartheta_j+\varsigma_j,$$

where

$$\begin{aligned}\vartheta_j&=\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\epsilon_{ij}A''(\widehat{\theta}_{ij})^{-1}}{\widehat{\sigma}_j},\\ \varsigma_j&=\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp-\boldsymbol{I}_p)\boldsymbol{e}_j}{\widehat{\sigma}_j}+\sqrt{n}\frac{T_{2j}+T_{3j}+T_{4j}}{\widehat{\sigma}_j}.\end{aligned}$$

For the first component, note that ϑ_j for $j=1,\ldots,p$ are independent conditional on $\{(\boldsymbol{x}_i,\boldsymbol{z}_i^*)\}_{i=1}^n$ and $\mathcal{D}_2$. Furthermore, $\vartheta_j\xrightarrow{\mathrm{d}}\mathcal{N}(0,1)$ for $j\in\mathcal{N}_p$ from Lemma A.5.2 by noting that $\widehat{\sigma}_j$ is also consistent to the conditional variance of ϑ_j. For the second component, from the proof of Theorem 6 and Lemma A.5.3, we know that

$$\max_{1\leq j\leq p}\sqrt{n}\frac{T_{2j}+T_{3j}+T_{4j}}{\widehat{\sigma}_j}=o_{\mathbb{P}}(1).$$

On the other hand, from Theorem 3 and Assumption 3, we also have

$$\begin{aligned}&\max_{1\leq j\leq p}\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp-\boldsymbol{I}_p)\boldsymbol{e}_j}{\widehat{\sigma}_j}\\ =&\max_{1\leq j\leq p}\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp-\mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp)\boldsymbol{e}_j}{\widehat{\sigma}_j}\\ &+\max_{1\leq j\leq p}\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{e}_j}{\widehat{\sigma}_j}\\ =&\mathcal{O}_{\mathbb{P}}\left(\sqrt{p}\max_{1\leq j\leq p}\|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp-\mathcal{P}_{\boldsymbol{\Gamma}^*}^\perp)\boldsymbol{e}_j\|_2\right)\\ &+\max_{1\leq j\leq p}\left\|\sqrt{n}\frac{\widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\boldsymbol{\Gamma}^*}{\widehat{\sigma}_j}\right\|_2\|(\boldsymbol{\Gamma}^{*\top}\boldsymbol{\Gamma}^*)^{-1}\|_{\mathrm{op}}\|\boldsymbol{\gamma}_j^*\|_2\\ =&\mathcal{O}_{\mathbb{P}}(p^{-\frac{1}{2}})+\mathcal{O}_{\mathbb{P}}(\sqrt{rp}\cdot p^{-1}\cdot 1)\\ =&o_{\mathbb{P}}(1).\end{aligned}$$

where we use the subexponential concentration of ϵ_{ij} conditional on $\{(\boldsymbol{x}_i, \boldsymbol{z}_i^*)\}_{i=1}^n$ and $\mathcal{D}_2$. Therefore, we have that $\max_{1\le j\le p} |\varsigma_j| = o_{\mathbb{P}}(1)$.

The rest of the proof follows similarly to the proof of Wang et al. [175, Theorem 3.4]. We present here for completeness.

Overall Type-I error control. Let $\varrho = |\mathcal{N}_p|^{-1} \sum_{j\in\mathcal{N}_p} \mathbb{1}\left(|t_j| > z_{\frac{\alpha}{2}}\right)$. To prove the overall Type-I error control, we will show the expectation of ϱ tends to α and its variance tends to zero. For the expectation, for any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}[\varrho] &= \frac{1}{|\mathcal{N}_p|} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|t_j| > z_{\frac{\alpha}{2}}\right) \\
&\le \frac{1}{|\mathcal{N}_p|} \sum_{j\in\mathcal{N}_p} \left[\mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(|\varsigma_j| > \epsilon\right)\right] \\
&= \frac{1}{|\mathcal{N}_p|} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \frac{1}{|\mathcal{N}_p|} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|\varsigma_j| > \epsilon\right) \\
&\le \frac{1}{|\mathcal{N}_p|} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(\max_{1\le j\le p} |\varsigma_j| > \epsilon\right) \\
&\to 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right),
\end{aligned}$$

where the last convergence is because the Cesaro mean converges to the same limit as $\lim_{n,p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) = 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right)$ and the second term $\mathbb{P}(\max_{1\le j\le p} |\varsigma_j| > \epsilon)$ varnishes. Similarly, we can also show that $\liminf_{n,p\to\infty} \mathbb{E}[\varrho] \ge 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right)$ for all $\epsilon > 0$. This implies that $\mathbb{E}[\varrho] \to \alpha$ as $n, p \to \infty$.

Next, we analyze the second moment. For any $\epsilon > 0$, the second moment can be upped bounded as

$$\begin{aligned}
\mathbb{E}[\varrho^2] &= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k\in\mathcal{N}_p} \mathbb{P}\left(|t_j| > z_{\frac{\alpha}{2}}, |t_k| > z_{\frac{\alpha}{2}}\right) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|t_j| > z_{\frac{\alpha}{2}}\right) + \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k\in\mathcal{N}_p, j\ne k} \mathbb{P}\left(|t_j| > z_{\frac{\alpha}{2}}, |t_k| > z_{\frac{\alpha}{2}}\right) \\
&\le \frac{1}{|\mathcal{N}_p|^2} \sum_{j\in\mathcal{N}_p} \mathbb{P}\left(|t_j| > z_{\frac{\alpha}{2}}\right) \\
&\quad + \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k\in\mathcal{N}_p, j\ne k} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon, |\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(|\varsigma_j| > \epsilon\right) + \mathbb{P}\left(|\varsigma_k| > \epsilon\right) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k\in\mathcal{N}_p, j\ne k} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon, |\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon\right) + o(1) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k\in\mathcal{N}_p, j\ne k} \mathbb{E}[\mathbb{P}(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon \mid \mathcal{C})\mathbb{P}(|\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon \mid \mathcal{C})] + o(1) \\
&\to \left[2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right)\right]^2,
\end{aligned}$$

where the last equality is from the independence of ϑ_j and ϑ_k condition on $\mathcal{C} = \{(\boldsymbol{x}_i, \boldsymbol{z}_i^*)\}_{i=1}^n \cup \mathcal{D}_2$. We can similarly obtain the lower bound. This implies that $\mathbb{E}[\varrho^2] \to \alpha^2$ and $\mathbb{V}(\varrho) \to 0$ as $n, p \to \infty$. Combining the previous results yields that $\varrho \xrightarrow{\mathrm{p}} \alpha$.

FWER control. To prove the second statement, note that

$$\begin{aligned}
\mathbb{P}\left(|\mathcal{N}_p|\varrho \geq 1\right) =& \mathbb{P}\left(\max_{j\in\mathcal{N}_p} |t_j| > \Phi^{-1}(1-\alpha/(2p))\right) \\
=& \mathbb{P}\left(\max_{j\in\mathcal{N}_p} |\vartheta_j + \varsigma_j| > \Phi^{-1}(1-\alpha/(2p))\right) \\
\leq& \mathbb{P}\left(\max_{j\in\mathcal{N}_p} |\vartheta_j| > \Phi^{-1}(1-\alpha/(2p)) - \max_{j\in\mathcal{N}_p} |\varsigma_j|\right) \\
\leq& \mathbb{P}\left(\max_{1\leq j\leq p} |\vartheta_j| > \Phi^{-1}(1-\alpha/(2p)) - \max_{j\in\mathcal{N}_p} |\varsigma_j|\right),
\end{aligned}$$

which is asymptotically upper bounded by α, after applying Gaussian approximation [30, Lemma 2.3] and the valid control of Bonferroni correction for i.i.d. normal random variables, by noting that $\Phi^{-1}(1-\alpha/(2p)) \to \infty$ as $p \to \infty$, and the result that $\max_{j\in\mathcal{N}_p} |\varsigma_j| = o_{\mathbb{P}}(1)$.

□

A.5.3 Technical lemmas

Lemma A.5.1. Under the same conditions as in Theorem 6, suppose event $\mathcal{E}_1$ holds, then the solution to optimization problem (2.4.4) exists with probability at least $1 - 2(nd)^{-c}$.

Proof of Lemma A.5.1. Define the matrix $\boldsymbol{S} = \mathbb{E}[\widehat{\omega}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top]$. We next show that (1) $\boldsymbol{S}$ is invertible and (2) the j-th column $\boldsymbol{u}^*$ of $\boldsymbol{S}^{-1}$ is feasible for the constraints of the optimization problem (2.4.4) with high probability. We split the proof into two parts, as below.

Part (1) Because $C^{-1} \leq \widehat{\omega}_i \leq C$, we have $C^{-1}\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top] \preceq \boldsymbol{S} \preceq C\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top]$. On the other hand, note that $\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top] = \boldsymbol{\Sigma}_x \succeq \lambda_{\min}(\boldsymbol{\Sigma}_x)\boldsymbol{I}_d$. Thus, for any unit vector $\boldsymbol{a} \in \mathbb{R}^d$, we have $\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a} \geq C^{-1}\lambda_{\min}(\boldsymbol{\Sigma}_x) > 0$. This establishes claim (1).

Part (2) Let $\boldsymbol{u}^*$ be the j-th column of $\boldsymbol{S}^{-1}$. By definition, we have $\boldsymbol{S}\boldsymbol{u}^* = \boldsymbol{e}_j$. Conditional on $\mathcal{D}_2$, we have that $\widehat{\omega}_i \boldsymbol{u}^{*\top}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{e}_k$ for $i = 1, \ldots, n$ are independent random variables with mean δ_{jk}. Because $\widehat{\omega}_i$ is bounded, we further have that $\widehat{\omega}_i \boldsymbol{u}^{*\top}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{e}_k$'s are independent sub-exponential random variables. Applying Bernstein's inequality as in the proof of Lemma A.4.2, we have with probability at least $1 - (nd)^{-c}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n \widehat{\omega}_i \boldsymbol{u}^{*\top}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{e}_j \right\|_\infty \leq \lambda_n,$$

where $\lambda_n \asymp \sqrt{\log(nd)/n}$. This also holds after taking account of the randomness of $\mathcal{D}_2$.

On the other hand, because ω_i is bounded away from zero and infinity, and $\|\boldsymbol{\Sigma}_x\|_{\mathrm{op}} = \mathcal{O}(1)$, it follows that $\|\boldsymbol{S}\|_{\mathrm{op}} = \mathcal{O}(1)$ and $\|\boldsymbol{u}^*\|_2 = \mathcal{O}(1)$. By the sub-Gaussianity of $\boldsymbol{x}_i$, we also have $\|\boldsymbol{X}\boldsymbol{u}^*\|_\infty = \max_{1\leq i\leq n} |\boldsymbol{x}_i^\top \boldsymbol{u}^*| \leq \tau_n$, with probability at least $1 - n^{-c}$. The above shows that $\boldsymbol{u}^*$ is feasible for optimization problem (2.4.4), which establishes the claim (2).

Finally, taking the union bound over the two probabilistic events finishes the proof. □

Lemma A.5.2 (Asymptotic normality)**.** Under the conditions in Theorem 6, it holds that

$$\sqrt{n}\sum_{i=1}^{n}\sigma_j^{-1}\widehat{\boldsymbol{u}}^\top\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\widehat{\boldsymbol{v}}_i \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

where $\sigma_j^2 = n^{-1}\,\mathbb{V}\left(\boldsymbol{u}^\top\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{v}_i \mid \{\boldsymbol{x}_i,\boldsymbol{z}_i\}_{i=1}^n,\mathcal{D}_2\right)$.

Proof of Lemma A.5.2. Note that when conditioning on the natural parameters $\boldsymbol{\Theta}^*$, ϵ_{ij}'s are independent (ν,α)-sub-exponential random variable as shown in the proof of Theorem 2. Define $\xi_i := \sigma_j^{-1}\widehat{\boldsymbol{u}}^\top\boldsymbol{x}_i\boldsymbol{\epsilon}_i\widehat{\boldsymbol{v}}_i$ for $i\in[n]$. Then ξ_i's are independent random variables with mean $\mathbb{E}[\xi_i \mid \mathcal{D}_2,\boldsymbol{X}] = 0$ and variance $\mathbb{V}(\xi_i\mid\mathcal{D}_2,\boldsymbol{X}) = 1$. It suffices to check the bounded variance condition and Lindeberg's condition.

Part (1) Boundedness of σ_j. We first show the boundedness of the variance

$$\sigma_j^2 = \widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i^2(\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\widehat{\boldsymbol{u}}.$$

Because $A''(\theta)\geq\kappa_1>0$ for all $\theta\in\mathcal{R}$, the quadratic term satisfies that

$$\boldsymbol{e}_j\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j\geq 0,$$

with equality holds if and only if $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j = \boldsymbol{0}_p$. On the other hand, we have

$$\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j\|_2 = \|\boldsymbol{e}_j - \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j\|_2 \geq 1 - \|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j\|_2 \gtrsim \Omega(1-p^{-1/2}),$$

where the last inequality is because

$$\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j\|_2 \leq \|\mathcal{P}_{\boldsymbol{\Gamma}^*}\boldsymbol{e}_j\|_2 + \|(\mathcal{P}_{\boldsymbol{\Gamma}^*}-\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}})\boldsymbol{e}_j\|_2 \lesssim \frac{1}{\sqrt{p}}$$

from Assumption 3 and Theorem 3. This implies that, when n,p are sufficiently large,

$$c^{-1}\leq\widehat{\omega}_i(\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\boldsymbol{e}_j)\leq c$$

for some constant $c>1$, under event $\mathcal{E}_1$. Thus, it is equivalent to show the boundedness of $\widehat{\sigma}_j^2 = \widehat{\boldsymbol{u}}^\top\widehat{\boldsymbol{S}}\widehat{\boldsymbol{u}}$, where $\widehat{\boldsymbol{S}} = n^{-1}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top$. From Lemma A.5.1, we know that $\boldsymbol{S} = \mathbb{E}[\widehat{\boldsymbol{S}}]$ has a bounded spectrum with high probability. The upper bound that $\widehat{\sigma}_j^2\leq\boldsymbol{S}_{jj}$ with high probability then follows by the sub-exponential concentration results as in the proof of Lemma A.5.1.

Next, we proceed to show the lower bound. Because $\widehat{\boldsymbol{u}}$ satisfies the constraint $|\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\widehat{\boldsymbol{u}}-1|\leq\lambda_n$, we have that $\sigma_j^2\geq\widehat{\boldsymbol{u}}^\top\widehat{\boldsymbol{S}}\widehat{\boldsymbol{u}}+t((1-\lambda_n)-\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\widehat{\boldsymbol{u}})$ for any $t>0$. Note that $\min_{\boldsymbol{v}\in\mathbb{R}^d}\boldsymbol{v}^\top\widehat{\boldsymbol{S}}\boldsymbol{v}+t((1-\lambda_n)-\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\boldsymbol{v}) = -t^2\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\boldsymbol{e}_j/4+t(1-\lambda_n)$ where the minimum is obtained when $\widehat{\boldsymbol{S}}\boldsymbol{v} = t\widehat{\boldsymbol{S}}\boldsymbol{e}_j/2$. We further have $\widehat{\sigma}_j^2\geq\max_{t\geq0}-t^2\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\boldsymbol{e}_j/4+t(1-\lambda_n)\geq(1-\lambda_n)^2/(\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\boldsymbol{e}_j)$. By the sub-Gaussianity of $\boldsymbol{x}_i$, $\boldsymbol{e}_j^\top\widehat{\boldsymbol{S}}\boldsymbol{e}_j\geq\boldsymbol{S}_{jj}+o_{\mathbb{P}}(1)$. We then have $\widehat{\sigma}_j^2\geq0.5/\boldsymbol{S}_{jj}$ when n and p are large enough.

Part (2) Lindeberg's condition. On the other hand, because

$$\max_{1\le i\le n}|\xi_i| \le \max_{1\le i\le n}|\widehat{\boldsymbol{u}}^\top \boldsymbol{x}_i|\|\boldsymbol{\epsilon}_i\|_2|\sigma_j^{-1}|\|\widehat{\boldsymbol{v}}_i\|_2 \lesssim \sqrt{n},$$

with probability at least $1-2(nd)^{-c}$, the Lindeberg's condition holds that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i^2\,\mathbb{1}\{|\xi_i|\ge\epsilon\sqrt{n}\}]=0$$

for all $\epsilon>0$. Applying Lindeberg's central limit theorem yields that

$$\sqrt{n}\sum_{i=1}^n \sigma_j^{-1}\widehat{\boldsymbol{u}}^\top\boldsymbol{x}_i\boldsymbol{\epsilon}_i\widehat{\boldsymbol{v}}_i \xrightarrow{\mathrm{d}}\mathcal{N}(0,1),$$

which finishes the proof. □

Lemma A.5.3 (Consistent estimators of σ_j)**.** Under conditions in Theorem 5 and condition (i) in Theorem 6, $\widehat{\omega}_j = A''(\widehat{\theta}_{ij})$ satisfies condition (ii) of Theorem 6. Furthermore, for the variance estimate defined in (2.4.6) using sample splitting procedure Algorithm A.5.5, it holds that $\widehat{\sigma}_j \xrightarrow{\mathrm{p}} \sigma_j$.

Proof of Lemma A.5.3. The boundedness of $\widehat{\omega}_j$ follows from (A.2.1).

By the sample splitting procedure Algorithm A.5.5, we know that for a given response $j\in[p]$, $\boldsymbol{Y}_j - A'(\boldsymbol{\Theta}_j^*)\in\mathbb{R}^n$ is independent of $\widehat{\boldsymbol{u}}$, $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{Z}}$, when conditioning on $\boldsymbol{X}$. However, noted that $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{Z}}$ may be specific to each $j\in I$. For the sake of simplicity, in the following proof, we will assume that $\boldsymbol{Y}-A'(\boldsymbol{\Theta}^*)$ is independent of $\widehat{\boldsymbol{u}}$, and a common set of estimators $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{Z}}$ when conditioning on $\boldsymbol{X}$; or equivalently $I=[p]$. Note that, however, the proof still works for the cases in Algorithm A.5.5, except the constructed debiased estimators only use responses in the index set I; namely, $\widehat{b}_{j1}^{\mathrm{de}} = \widehat{b}_{j1} + \widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i([\boldsymbol{y}_i]_I - [A'(\widehat{\boldsymbol{\theta}}_i)]_I)^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}_I}^{\perp}\boldsymbol{e}_j$.

Recall that

$$\sigma_j^2 = \widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i^2(\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j)\boldsymbol{x}_i\boldsymbol{x}_i^\top\widehat{\boldsymbol{u}},$$

and

$$\widehat{\sigma}_j^2 = \widehat{\boldsymbol{u}}^\top\frac{1}{n}\sum_{i=1}^n\widehat{\omega}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\widehat{\boldsymbol{u}}.$$

Let $a_i = \widehat{\omega}_i^2(\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j)$ and $a_i' = \widehat{\omega}_i$. We begin by bounding the difference between the two:

$$\begin{aligned}
|\widehat{\sigma}_j^2-\sigma_j^2| &= \left|\frac{1}{n}\sum_{i=1}^n(a_i'-a_i)\cdot(\widehat{\boldsymbol{u}}\boldsymbol{x}_i)^2\right| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{u}}\boldsymbol{x}_i)^4}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n(a_i'-a_i)^2} \\
&\lesssim \tau_n^2\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n(a_i'-a_i)^2},
\end{aligned}$$

where the first inequality is from Holder's inequality and the second inequality is due to the second constraint of the optimization problem (2.4.4). For the second factor above, note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(a_i'-a_i)^2\\
=&\max_{i\in[n]}\widehat{\omega}_i^4\cdot\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j-A''(\widehat{\boldsymbol{\theta}}_i))^2. \qquad \text{(A.5.4)}
\end{aligned}$$

Each term inside the square can be decomposed into

$$\begin{aligned}
&\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j-A''(\widehat{\theta}_{ij})\\
=&\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathrm{diag}(A''(\boldsymbol{\theta}_i^*)-A''(\widehat{\boldsymbol{\theta}}_i))\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j\\
&\quad-[2\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}A''(\widehat{\boldsymbol{\theta}}_i)\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j+\boldsymbol{e}_j^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}A''(\widehat{\boldsymbol{\theta}}_i)\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j]\\
=&:T_1+T_2. \qquad \text{(A.5.5)}
\end{aligned}$$

Now note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}T_1^2\lesssim&\frac{1}{n}\sum_{i=1}^{n}(A''(\boldsymbol{\theta}_{ij}^*)-A''(\widehat{\boldsymbol{\theta}}_{ij}))^2\\
=&\frac{1}{n}(\widehat{\boldsymbol{\Theta}}_j-\boldsymbol{\Theta}_j^*)^\top\mathrm{diag}(A'''(\boldsymbol{\Theta}_j'))(\widehat{\boldsymbol{\Theta}}_j-\boldsymbol{\Theta}_j^*)\\
\lesssim&\frac{1}{n}\|\widehat{\boldsymbol{\Theta}}_j-\boldsymbol{\Theta}_j^*\|_2^2, \qquad \text{(A.5.6)}
\end{aligned}$$

where the first inequality is due to the boundedness of A'' on $\mathcal{R}$, and the bounded spectral of the projection matrix $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}$, and noting that $\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j\|_2\lesssim\mathcal{O}_{\mathbb{P}}(p^{-1/2})$; the second equality is from Taylor expansion with $\boldsymbol{\Theta}_j'=(\theta_{1j}',\ldots,\theta_{nj}')$ and θ_{ij}' being between $\widehat{\theta}_{ij}$ and θ_{ij}^* for $i=1,\ldots,n$; and the second inequality is from the continuity and boundedness of A''' on $\mathcal{R}_C$.

On the other hand,

$$\frac{1}{n}\sum_{i=1}^{n}T_2^2\lesssim p^{-1} \qquad \text{(A.5.7)}$$

by noting that $\|\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{e}_j\|_2\lesssim\mathcal{O}_{\mathbb{P}}(p^{-1/2})$ again.

By applying triangle inequality on (A.5.4) and combining (A.5.5)-(A.5.7), we further have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(a_i'-a_i)^2\\
\lesssim&\max_{i\in[n]}\widehat{\omega}_i^4\cdot\left(\frac{1}{n}\sum_{i=1}^{n}T_1^2+\frac{1}{n}\sum_{i=1}^{n}T_2^2\right)\\
\lesssim&\frac{1}{n}\|\widehat{\boldsymbol{\Theta}}_j-\boldsymbol{\Theta}_j^*\|_2^2,
\end{aligned}$$

Algorithm A.5.5 Data splitting procedure.

Input: Data $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^d \times \mathbb{R}^p$ for $i = 1, \ldots, 2n$.

1: Split the full data into two disjoint datasets $\mathcal{D}_1 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1, \ldots, n\}$ and $\mathcal{D}_2 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = n+1, \ldots, 2n\}$.
2: Apply Algorithm 1 on $\mathcal{D}_2$ to obtain the estimates $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Gamma}}$.
3: **for** $j = 1, 2, \ldots, p$ **do**
4: Select a subset $I \subseteq [p] \cap \{j\}$ and set $I^c = [p] \setminus I$.
5: Based on $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Gamma}}$, use partial data $(\boldsymbol{X}, \boldsymbol{Y}_{I^c})$ to estimate $\widehat{\boldsymbol{Z}}$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots \boldsymbol{x}_n]^\top$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots \boldsymbol{y}_n]^\top$ and $\boldsymbol{Y}_{I^c} = [\boldsymbol{Y}_\ell]_{\ell \in I^c}$.
(Alternatively, Step 2-5 can be combined such that $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{Z}}$ are estimated jointly for gene j.)
6: Based on $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{Z}}$, estimate $\widehat{\omega}_i$'s and $\widehat{\boldsymbol{u}}$ on $(\boldsymbol{X}, \boldsymbol{Y}_I)$.
7: Calculate the test statistics z_j for gene j.
8: **end for**

Output: A set of test statistics $\{z_j \mid j = 1, \ldots, p\}$.

where in the last inequality, we also use the boundedness of $\widehat{\omega}_i$. This implies that

$$\begin{aligned}
|\widehat{\sigma}_j^2 - \sigma_j^2| &\lesssim \tau_n^2 \sqrt{\frac{1}{n}\|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^*\|_2^2} \\
&\lesssim \tau_n^2 \sqrt{\frac{1}{n}\|\boldsymbol{X}(\widehat{\boldsymbol{b}}_j - \boldsymbol{b}_j^*)\|_2^2 + \frac{1}{n}\|\widehat{\boldsymbol{E}}_j - \boldsymbol{E}_j^*\|_2^2} \\
&\lesssim \tau_n^2 r_{n,p} \\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

where $r_{n,p}$ is defined in Theorem 5. Here the concentration of $\|\boldsymbol{X}(\widehat{\boldsymbol{b}}_j - \boldsymbol{b}_j^*)\|_2^2$ is from Theorem 5 and the one of $\|\widehat{\boldsymbol{E}}_j - \boldsymbol{E}_j^*\|_2^2$ is from (A.4.24) as in the proof of Lemma A.4.3. This implies that $\widehat{\sigma}_j^2 \xrightarrow{\mathrm{p}} \sigma_j^2$. The conclusion then follows by applying the continuous mapping theorem. □

A.6 Computational aspects

A.6.1 Exponential family

Some commonly used exponential families, the exact formulas of the log-partition functions and other statistics, are summarized in Table A.61.

Distribution	**Extra parameter**	**Base measure** $h(y)$	**Sufficient statistics** $T(y)$	**Domain** $\mathrm{dom}(A(\theta))$	**Log-partition** $A(\theta)$	**Mean** $\mu = A'(\theta)$	**Variance** $A''(\theta)$
Gaussian	variance σ^2	$\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$\frac{y}{\sigma}$	$\mathbb{R}$	$\frac{\theta^2}{2}$	θ	1
Bernoulli		1	y	$\mathbb{R}$	$\log(1+e^\theta)$	$\frac{1}{1+e^{-\theta}}$	$\mu(1-\mu)$
Binomial	number of trials m	$\binom{m}{y}$	y	$\mathbb{R}$	$m\log(1+e^\theta)$	$\frac{m}{1+e^{-\theta}}$	$\mu\left(1-\frac{\mu}{m}\right)$
Poisson		$\frac{1}{y!}$	y	$\mathbb{R}$	e^θ	e^θ	e^θ
Negative Binomial	number of failures ϕ	$\binom{y+\phi-1}{y}$	y	$\mathbb{R}_-$	$-\phi\log(1-e^\theta)$	$\phi\frac{e^\theta}{1-e^\theta}$	$\phi\frac{e^\theta}{(1-e^\theta)^2}$

Table A.61: Summary of exponential family in canonical form.

A.6.2 Optimization details

Initialization. Our initialization procedure for optimalization problem (2.3.4) is inspired by Lin et al. [105].

- Initialize the marginal effects $\boldsymbol{F}$ by solving a generalized linear model without considering the latent variables. When the fitting of GLM is numerically unstable, one can also add a small ridge penalty $\lambda = 10^{-5}$.
- Initialize $\boldsymbol{W}$ and $\boldsymbol{\Gamma}$ using the SVD of the matrix $\log(\boldsymbol{Y}+1) = \boldsymbol{U}_Y\boldsymbol{\Sigma}_Y\boldsymbol{V}_Y^\top$ for Poisson likelihood or Negative Binomial likelihood with log link. Let $\boldsymbol{W} = (\mathcal{P}_{\boldsymbol{X}}^\perp\boldsymbol{U}_Y\boldsymbol{\Sigma}_Y^{1/2})_{1:r}$ and $\boldsymbol{\Gamma} = (\boldsymbol{V}_Y\boldsymbol{\Sigma}_Y^{1/2})_{1:r}$ be the first r columns of the corresponding matrices. Here the projection $\mathcal{P}_{\boldsymbol{X}}^\perp$ ensures that $\boldsymbol{W}$ is uncorrelated with $\boldsymbol{X}$. In particular, when the intercept is included in the covariates, the initial value of $\boldsymbol{W}$ also has zero means per column.

To initialize variables for optimization problem (2.3.5):

- Initialize the direct effects $\boldsymbol{B}$ as $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^\perp\widehat{\boldsymbol{F}}$.
- Initialize $\boldsymbol{Z}$ and $\boldsymbol{\Gamma}$ using the SVD of the matrix $\boldsymbol{X}\widehat{\boldsymbol{F}}^\top\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} + \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Gamma}}^\top = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^\top$. Let $\boldsymbol{Z} = (\boldsymbol{U}'\boldsymbol{\Sigma}'^{1/2})_{1:r}$ and $\boldsymbol{\Gamma} = (\boldsymbol{V}'\boldsymbol{\Sigma}'^{1/2})_{1:r}$. Because the latter has the same column space as $\widehat{\boldsymbol{\Gamma}}$, we simply treat the latter as $\widehat{\boldsymbol{\Gamma}}$ in optimization problem (2.3.5) with a light abuse of notation.

Alternative maximization The alternative maximization Algorithm A.6.6 is used to perform nonconvex matrix factorization. In our setup where the objective function is convex in the natural parameter, each iteration of Algorithm A.6.6 is simply solving two convex optimization subproblems.

Algorithm A.6.6 Joint maximum likelihood estimation by alternative maximization

Input: Data $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ from exponential family with log-partition function A, the regularization parameter λ, and initial value $\boldsymbol{l}_i^{(0)} \in \mathcal{D}_{l_i}$, $\boldsymbol{r}_j^{(0)} \in \mathcal{D}_{r_j}$ for $i \in [n]$ and $j \in [p]$.

Initialize the iteration number $t = 0$.
while not converged **do**
  $t \leftarrow t + 1$.
  for $i = 1, 2, \ldots, n$ **do**

$$\boldsymbol{l}_i^{(t)} \in \operatorname*{argmax}_{\boldsymbol{l}\in\mathcal{D}_{l_i}} \frac{1}{p}\sum_{j=1}^{p}\left(y_{ij}\boldsymbol{l}^\top \boldsymbol{r}_j^{(t-1)} - A(\boldsymbol{l}^\top \boldsymbol{r}_j^{(t-1)})\right)$$

  end for
  for $j = 1, 2, \ldots, p$ **do**

$$\boldsymbol{r}_j^{(t)} \in \operatorname*{argmax}_{\boldsymbol{r}\in\mathcal{D}_{r_j}} \frac{1}{n}\sum_{i=1}^{n}\left(y_{ij}\boldsymbol{l}_i^{(t)\top} \boldsymbol{r} - A(\boldsymbol{l}_i^{(t)\top} \boldsymbol{r})\right) - \lambda\|\boldsymbol{r}\|_1$$

  end for
end while

Output: $\boldsymbol{L} = [\boldsymbol{l}_i^{(t)}]_{i\in[n]}^\top$ and $\boldsymbol{R} = [\boldsymbol{r}_j^{(t)}]_{j\in[p]}^\top$ with $\boldsymbol{L}\boldsymbol{R}^\top$ being the estimated natural parameters.

By default, we use the inexact line search algorithm with an initial step size of 0.1 and a shrinkage factor of 0.5 for each iteration. The search is stopped if the Armijo rule is satisfied with tolerance 10^{-4} or the number of iterations reaches 20. We early stop the alternative maximization if the objective value does not increase more than a tolerance of 10^{-4} for 20 iterations.

Estimation of dispersion parameters To estimate the dispersion parameter, we first fit GLMs on the data and obtain the estimated mean expression of gene j, denoted as $\widehat{\mu}_j$ for $j = 1, \ldots, p$. Note that when y_{ij} comes from a Negative Binomial distribution, its variance is given by

$$\mathbb{V}(y_{ij} \mid \theta_{ij}) = \mu\,(1 + \alpha_j\mu)$$

where $\mu = \mathbb{E}[y_{ij} \mid \theta_{ij}]$ is the conditional mean while α_j is the dispersion parameter of the NB1 form. In the form of exponential family in Table A.61 parameterized by the parameter ϕ_j, α_j is the reciprocal of ϕ_j, namely, $\alpha_j = 1/\phi_j$. By methods of moments, we can solve the following equation to obtain an estimator $\widehat{\phi}_j$ for ϕ_j:

$$\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \widehat{\mu}_j)^2 = \widehat{\mu}_j\,(1 + \alpha\widehat{\mu}_j)\,.$$

Finally, we clip $\widehat{\alpha}_j$ to be in $[10^{-2}, 10^2]$ and set $\widehat{\phi}_j = 1/\widehat{\alpha}_j$.

A.6.3 Choice of hyperparameters in practice

The main text provides theoretical results for the proposed method under certain assumptions. The proposed algorithm Algorithm 1 requires the choice of hyperparameters, such as the rank r, the regularization parameters λ in optimization problem (2.3.5), and (τ_n, λ_n) in optimization

problem (2.4.4). Although the theoretical orders of some parameters are provided for consistency and asymptotic normality, the choice of hyperparameters in practice is crucial for the performance of the proposed method. Below, we discuss the choice of hyperparameters in practice.

Boundedness constant C. The boundedness constant C is a reasonably large constant that ensures a finite solution to optimization problems exists. In our simulations, estimating the model parameters is not sensitive to the choice of boundedness constant as long as it is set to be sufficiently large; see also Chen and Li [26, Appendix D] for detailed discussions. Therefore, in our implementation, instead of restricting the parameters to be bounded, we project the gradient at each step of the alternative maximization onto the L_2-norm ball with radius $2C'$ for some constant C'. A smaller value of C' equals decreasing the learning rates while improving the numerical stability. We set C' to be 10^5 and 10^3 for experiments with Poisson and Negative Binomial likelihoods, respectively.

Lasso penalty λ. For the lasso penalty $\lambda = c_1\sqrt{\log p/n}$, one can use cross-validation to tune the lasso penalty for optimal log-likelihood. However, because the estimation results are insensitive to the choice of this penalty, we simply set c_1 to be 0.02 and 0.01 for experiments with Poisson and Negative Binomial likelihoods, respectively.

The number of factors r. For the number of factors r, the joint-likelihood-based information criterion (JIC) proposed by Chen and Li [26] can be utilized to select a proper number of latent factors. The JIC value is the sum of deviance and a penalty on model complexity:

$$\begin{aligned}\mathrm{JIC}(\widehat{\boldsymbol{\Theta}}^{(r)}) &= \text{deviance} + \nu(n, p, d + r)\\ &= -2\sum_{i\in[n],j\in[p]} \log p(y_{ij} \mid \widehat{\theta}_{ij}^{(r)}) + c_{\mathrm{JIC}} \cdot \frac{(d+r)\log(n \wedge p)}{n \wedge p},\end{aligned}$$

where $\widehat{\boldsymbol{\Theta}}^{(r)}$ is the estimated natural parameter matrix with r latent factors and d observed covariates, and $c_{\mathrm{JIC}} > 0$ is a universal constant set to be one in all our simulations. As shown by Chen and Li [26], minimizing the empirical JIC yields a consistent estimate for the number of factors in generalized linear factor models. As an illustration, we compute the values of JIC at different numbers of factors on simulated datasets and visualize them in Figure A.61. When the unmeasured confounding effects are strong, the default choice of $c_{\mathrm{JIC}} = 1$ gives reasonable estimates for the number of factors under both Poisson and Negative Binomial likelihoods. Because the complexity term is a linear function in r, one can also inspect the increment of log-likelihood compared to the increment of the complexity term as a function of r, as shown in the right panel of Figure A.61. For real-world datasets, this can help to adjust the penalty level c_{JIC} to select a suitable value of r to achieve a sufficient reduction of negative log-likelihood while avoiding overfitting.

Debiasing parameters (τ_n, λ_n). For inference, two parameters (τ_n, λ_n) are to be specified. However, the parameter τ_n is less important because as long as both the covariate $\boldsymbol{x}_i$ and the projection direction $\boldsymbol{u}$ are bounded in L_∞-norm, the second constraint in Equation (4.4) will always be satisfied. For this reason, we can ignore the second constraint and solve the relaxed optimization problem, similar to the implementation of Cai et al. [21]. Therefore, one only needs to determine the parameter $\lambda_n = c_2\sqrt{\log(n)/n}$.

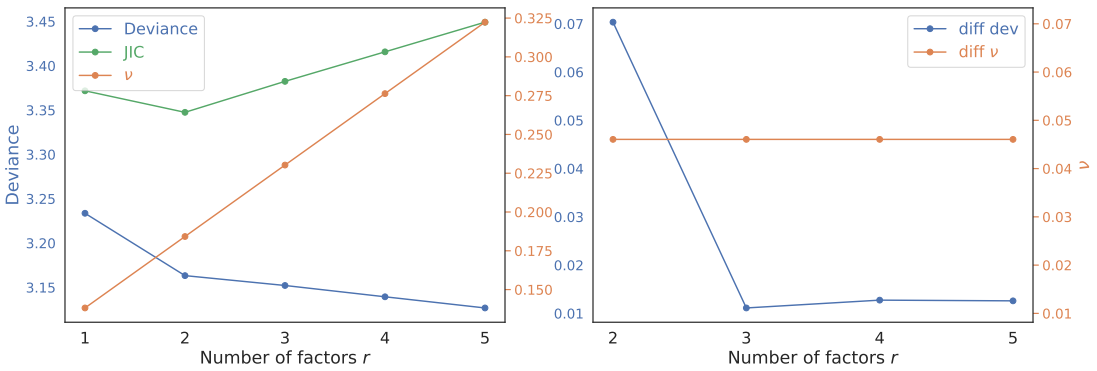

Figure A.61: The left panel shows the deviance and the complexity penalty ν at different numbers of factors r. The JIC is the sum of deviance and ν. The right panel shows the decrement of the deviance and the complexity at different numbers of factors r. The values are computed from one simulation in Section 2.5 with $n = 100$ and $r^* = 2$ underlying factors.

To address this, Cai et al. [21] only implemented a single value for c_2. However, we propose a more effective heuristic method to guide the selection of c_2. Specifically, we enumerate different values of $c_2 \in \{0.001, 0.002, \ldots, 0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1\}$ and compute the median and median absolute deviation (MAD) of the corresponding empirical z-statistics. We then generate the scree plot of the two summarized statistics. As shown in Figure A.62, as λ_n increases, both the median and the MAD of the empirical null distribution change. Specifically, as λ_n increases, the median decreases, while the MAD increases and then decreases. Therefore, when λ_n is too small, the empirical null distribution concentrates around 0, and the resulting tests will be conservative. On the other hand, when λ_n is too large, the tests will be anti-conservative. Therefore, a reasonable choice for λ_n is such that the absolute value of the median is not too large while the MAD of the corresponding test statistics is near one.

For simulations with Poisson likelihood, according to the scree plot, the adaptive choice of the value c_2 would be the largest value that makes the median deviate from 0 by no more than a threshold of 0.1. Analogously, we set the median deviation threshold to be 0.025 for the Negative Binomial simulations. Note that any value below the selected λ_n also provides valid inference results but with lower power.

A.6.4 Negative binomial likelihood with non-canonical link

While theoretically nice, the canonical link function for Negative Binomial distributions (NB-C) is not recommended in general because its natural parameter value is always negative, but linear predictors ought to be unbounded in general. Numerical instability may occur in the boundary of the natural parameters. Furthermore, the NB-C model is sensitive to the initial values and may converge to a local solution.

The common choice of a link function for generalized linear models with Negative Binomial likelihood is the log link [3]. Below, we show how to incorporate non-canonical ink into our framework.

For Negative Binomial distribution, recall that ϕ is a parameter that represents the number of failures as in Table A.61. Define the Negative Binomial canonical link A'^{-1} and log link L^{-1}, such that $A'(\theta) = \phi e^{\theta}/(1 - e^{\theta})$ and $L(\xi) = e^{\xi}$.

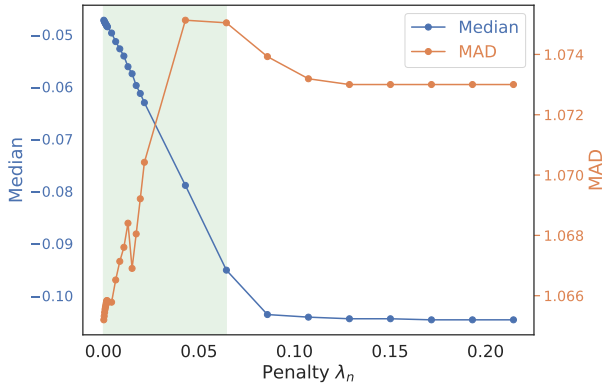

Figure A.62: The median and MAD of the z-statistics as a function of the regularization parameter λ_n computed from one simulation in Section 2.5.1 with $n = 100$ and $r = 2$. The shaded region indicates feasible values of λ_n, for which the absolute values of the medians of the corresponding test statistics are less than 0.1.

Let θ and ξ be the natural parameter and its representation under the log link; namely, the mean μ can be obtained from them through the corresponding link functions:

$$\mu = A'(\theta) = L(\xi).$$

This gives rise to the transformation equations:

$$e^{\xi} = \phi \frac{e^{\theta}}{1 - e^{\theta}}, \qquad e^{\theta} = \frac{e^{\xi}}{\phi + e^{\xi}},$$

and

$$\theta = A'^{-1}(L(\xi)) = \log \frac{e^{\xi}}{\phi + e^{\xi}}. \tag{A.6.1}$$

Note that the negative log-likelihood is given by

$$\begin{aligned} l(\xi) &:= -y \cdot (A')^{-1}(L(\xi)) + A((A')^{-1}(L(\xi))) \\ &= -y \log \frac{e^{\xi}}{\phi + e^{\xi}} - \phi \log \frac{\phi}{\phi + e^{\xi}}, \end{aligned}$$

which has gradient and hessian:

$$\frac{\partial l}{\partial \xi} = \frac{\partial l}{\partial \theta} \frac{\partial \theta}{\partial \xi} = -(y - A'(\theta)) \frac{\phi}{\phi + e^{\xi}} \tag{A.6.2}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \xi^2} &= \frac{\partial^2 l}{\partial \theta^2} \frac{\partial \theta}{\partial \xi} + \frac{\partial l}{\partial \theta} \frac{\partial^2 \theta}{\partial \xi^2} \\ &= A''(\theta) \left(\frac{\partial \theta}{\partial \xi} \right)^2 + \frac{\partial l}{\partial \theta} \frac{\partial^2 \theta}{\partial \xi^2} \\ &\approx \frac{\phi e^{\xi}}{\phi + e^{\xi}} \end{aligned} \tag{A.6.3}$$

where the last line is because the conditional expectation on $y - A'(\theta)$ given θ is zero, $\mathbb{E}[\partial l/\partial\theta \mid \theta] = 0$, so that the second term is ignorable. The latter approximation approach is also used in classic GLM to derive the asymptotic variance of the estimates.

For n i.i.d. samples $(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{y}_i)$, the linear predictor reads that $\boldsymbol{\Xi} = \boldsymbol{X}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top$ when using the log link. Based on the relationship (A.6.1), we can perform estimation and inference for the log link function, as described below.

For estimation, the objective function (2.2.3) now becomes:

$$l(\boldsymbol{\Xi}) = l(\boldsymbol{\Gamma}, \boldsymbol{Z}, \boldsymbol{B}) = -\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[p]}(y_{ij}A'^{-1}(L(\xi_{ij})) - L(\xi_{ij})).$$

Even though the new objective is now nonconvex in the parameter $\boldsymbol{\Xi}$, the alternative maximization algorithm Algorithm A.6.6 is still applicable to it, because the gradient can be computed based on (A.6.2). If we initialize $\widehat{\boldsymbol{F}}$ from GLM estimates and treat it as fixed, then solving optimization problem (2.3.2) reduces to a nonconvex matrix factorization problem. Under this setting, there is a rich literature on establishing the estimation error for $\boldsymbol{W}^*$ and $\boldsymbol{\Gamma}^*$ given that the initial value is close to the truth; see Lin et al. [105], Wang et al. [176] among the others. In other words, we may also obtain error bounds on $\|\boldsymbol{\Xi}^* - \widehat{\boldsymbol{\Xi}}\|_{\mathrm{F}}^2$ by imposing additional conditions.

For inference, we simply apply the chain rule and (A.6.2)-(A.6.3) to rewrite (2.4.1) as:

$$\widehat{b}_{j1}^{\mathrm{de}} = \widehat{b}_{j1} + \boldsymbol{u}^\top\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{y}_i - A'(\widehat{\boldsymbol{\theta}}_i))^\top\mathrm{diag}\left(\left\{\frac{\partial\theta}{\partial\xi}\Big|_{\xi=\widehat{\xi}_{ij}}\right\}_{j\in[p]}\right)\boldsymbol{v}_i,$$

with

$$\boldsymbol{v}_i = \omega_i\mathrm{diag}\left(\left\{\frac{\partial\theta}{\partial\xi}\Big|_{\xi=\widehat{\xi}_{ij}}\right\}_{j\in[p]}\right)^{-2}\mathrm{diag}(A''(\widehat{\boldsymbol{\theta}}_i))^{-1}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\boldsymbol{e}_j,$$

$$\omega_i = \mathbb{E}\left[\frac{\partial^2 l}{\partial\xi^2}\,\Big|\,\xi = \widehat{\xi}_{ij}\right] = A''(\widehat{\theta}_{ij})\left(\frac{\partial\theta}{\partial\xi}\Big|_{\xi=\widehat{\xi}_{ij}}\right)^2 = \frac{\phi e^{\xi}}{\phi + e^{\xi}}.$$

Because when $\mathcal{R}_C$ is bounded, the derivative function $\partial\theta/\partial\xi$ is Lipschitz continuous, the estimation error of it:

$$\sum_{j=1}^{p}(\partial\theta/\partial\xi|_{\xi=\widehat{\xi}_{ij}^*} - \partial\theta/\partial\xi|_{\xi=\widehat{\xi}_{ij}})^2 \lesssim \|\boldsymbol{\xi}_i^* - \widehat{\boldsymbol{\xi}}_i\|_2^2$$

can be bounded if $\boldsymbol{\xi}_i^*$ can be well estimated. Similarly, the estimation error of $\boldsymbol{\Theta}^*$ can be controlled because θ_{ij}^* is a Lipschitz continuous function of ξ_{ij}^*. Thus, Theorem 6 also applies if

$$\|\boldsymbol{\Xi}^* - \widehat{\boldsymbol{\Xi}}\|_{\mathrm{F}}^2 \lesssim \sqrt{n \vee p} \tag{A.6.4}$$

$$\max_{1\le j\le p}\frac{1}{\sqrt{n}}\|\boldsymbol{Z}^*\boldsymbol{\Gamma}_j^{*\top} - \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Gamma}}_j^\top\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{n \wedge p}} \tag{A.6.5}$$

$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{1,1} \lesssim \frac{\sqrt{sd}}{\sqrt{n \wedge p}} \tag{A.6.6}$$

$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{n \wedge p}}, \tag{A.6.7}$$

where (A.6.4) requires an analysis tool from nonconvex matrix factorization, (A.6.5) is a direct consequence from (A.6.4) similar to Corollary 4, and (A.6.6)-(A.6.7) requires non-asymptotic analysis as in the proof of Theorem 5 but for nonconvex objectives instead.

A.7 Extra experiment results

A.7.1 Efficiency loss of sample splitting

To evaluate the efficiency loss caused by sample splitting described in Algorithm A.5.5, we conduct the experiments with different splitting proportions and compare their results. To apply Algorithm A.5.5, we split the p genes into 2 groups with equal sizes, so that $I_1 = \{1, \ldots, p/2\}$ and $I_2 = \{p/2 + 1, \ldots, p\}$. For each of the groups I, the optimization is jointly conducted based on $\boldsymbol{X}$, $\boldsymbol{Y}_{I^c}$ and $\mathcal{D}_2$, and the inference is conducted for genes in I. As summarized in Table A.72, the performance on Type-I error and FDP control is similar across different splitting ratios. However, the power and precision are affected when the ratio of observations reserved for inference is too small. This suggests that one should leave more observations to conduct the debias step. Lastly, we see similar performance even without sample splitting, suggesting that the validity of the inferential procedure could be true even without sample splitting.

ratio split	type-I error	FDP	power	precision
0.2	0.050	0.200	0.454	0.610
0.4	0.049	0.193	0.755	0.920
0.6	0.050	0.191	0.901	1.000
0.8	0.051	0.195	0.963	1.000
no splitting	0.051	0.219	0.987	1.000

Table A.72: Performance with varying ratios of observations reserved for inference, under the same data setup in Section 2.5.1 with $n = 250$ and $r = 2$. The values are medians over 100 simulated datasets.

A.7.2 The blessing of dimensionality

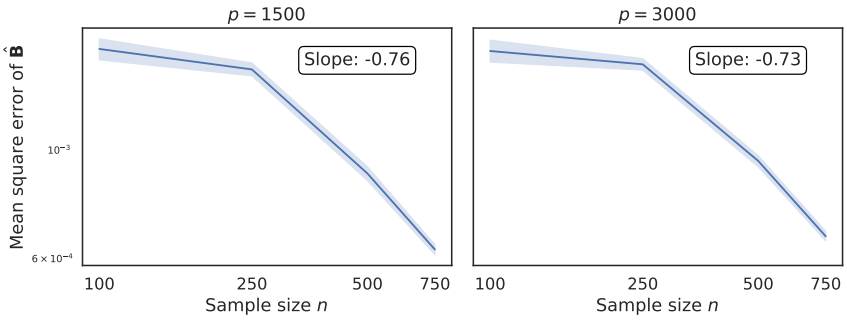

Figure A.73: The mean square error of $\widehat{\boldsymbol{B}}$ with varying outcome dimension p and sample size n, displayed on the log-log scale. When the outcome dimension p is sufficiently large (not growing exponentially in n), the estimation error of $\boldsymbol{B}$ is mainly driven by the sample size n. The slope is estimated using sample sizes larger than 100. The data generating process is given in Section 2.5.1.

A.7.3 Information about lupus data

Cell type	Number of samples n	Number of genes p	Proportion of non-zeros
T4	256	1255	0.398
cM	256	1208	0.434
B	254	1269	0.417
T8	256	1281	0.471
NK	256	1178	0.385

Table A.73: Summary statistics of the preprocessed lupus datasets in each cell type. The last column represents the proportion of non-zero count in the gene expression matrix.

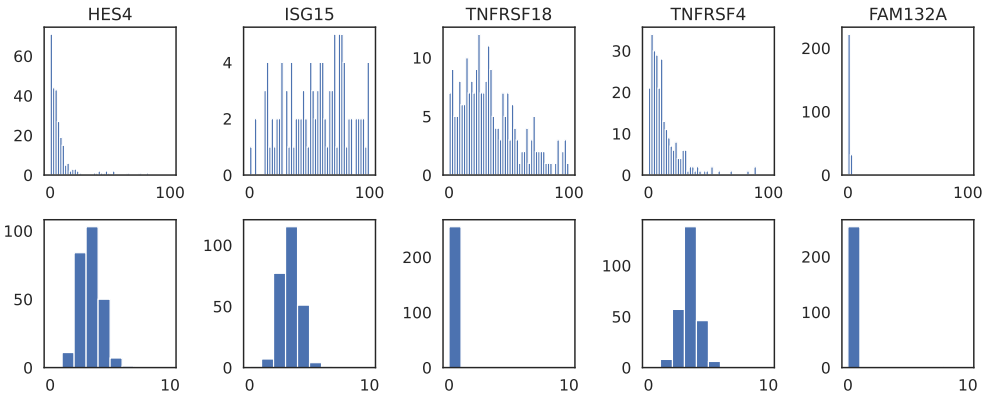

Figure A.74: Histograms of expressions of 5 genes on the T4 cell type. The first row shows the raw pseudo-bulk counts and the second row shows the counts after library size normalization and log1p transformation, which is used for CATE. Due to the sparsity of the gene expressions, some genes are not distributed like normal after transformation.

A.7.4 Extra results on lupus datasets

Sensitivity analysis for the number of latent factors

We inspect the sensitivity of GCATE-subset to the number of latent factors r. By utilizing JIC (2.3.1), we have selected $r = 7$ for the T4 cell type, which is close to the number of major covariates we drop. In Table A.74 and Table A.75, we examine the performance of GCATE-subset and GCATE-full for different values of r. Remarkably, the resulting distributions of z-statistics generated by GCATE, across varying numbers of factors r, are similar to the standard normal distribution when $r \geq 3$ because the MAD is close to one. Thus, JIC can serve as a valuable criterion for determining the appropriate number of latent factors for GCATE. Furthermore, it is noteworthy that the number of discoveries remains consistent when r falls within a reasonable range. These observations collectively suggest the stability of GCATE's inferential outcomes within this range of reasonable factor selections.

r	mean	median	mad	num_sig	deviance	JIC
1	0.348	-0.018	1.551	302	3.578	3.600
2	0.361	-0.040	1.539	313	3.565	3.592
3	0.141	-0.001	1.145	57	3.553	3.586
4	0.140	0.063	1.087	37	3.546	3.584
5	0.130	0.043	1.067	33	3.539	3.582
6	0.140	0.051	1.084	39	3.532	3.581
7	0.128	0.057	1.033	22	3.526	3.580
8	0.139	0.069	1.044	18	3.521	3.580
9	0.143	0.073	1.032	18	3.516	3.581
10	0.155	0.095	1.038	20	3.513	3.583

Table A.74: The summary of the z-statistics and model fitness for a varying number of latent factors r for GCATE-subset analysis. The metrics include the mean, median, median absolute deviation (mad), and the total number of significant genes of q-value less than 0.2. The last two columns show the deviance (2 times the negative log-likelihood) and the JIC model selection criteria (2.3.1) with $c_{\text{JIC}} = 0.25$.

r	mean	median	mad	num_sig	deviance	JIC
1	0.242	-0.014	1.350	200	3.552	3.650
2	0.163	0.013	1.192	93	3.542	3.650
3	0.108	0.010	1.111	21	3.534	3.653
4	0.143	0.052	1.119	20	3.528	3.658
5	0.151	0.066	1.071	23	3.523	3.664
6	0.165	0.071	1.119	24	3.518	3.670
7	0.174	0.077	1.104	29	3.514	3.676
8	0.170	0.092	1.067	25	3.510	3.683
9	0.170	0.100	1.070	38	3.507	3.691
10	0.178	0.103	1.075	38	3.502	3.697

Table A.75: The summary of the z-statistics and model fitness for a varying number of latent factors r for GCATE-subset analysis. The metrics include the mean, median, median absolute deviation (mad), and the total number of significant genes of q-value less than 0.2. The last two columns show the deviance (2 times the negative log-likelihood) and the JIC model selection criteria (2.3.1) with $c_{\text{JIC}} = 0.5$.

Selection of hyperparameters

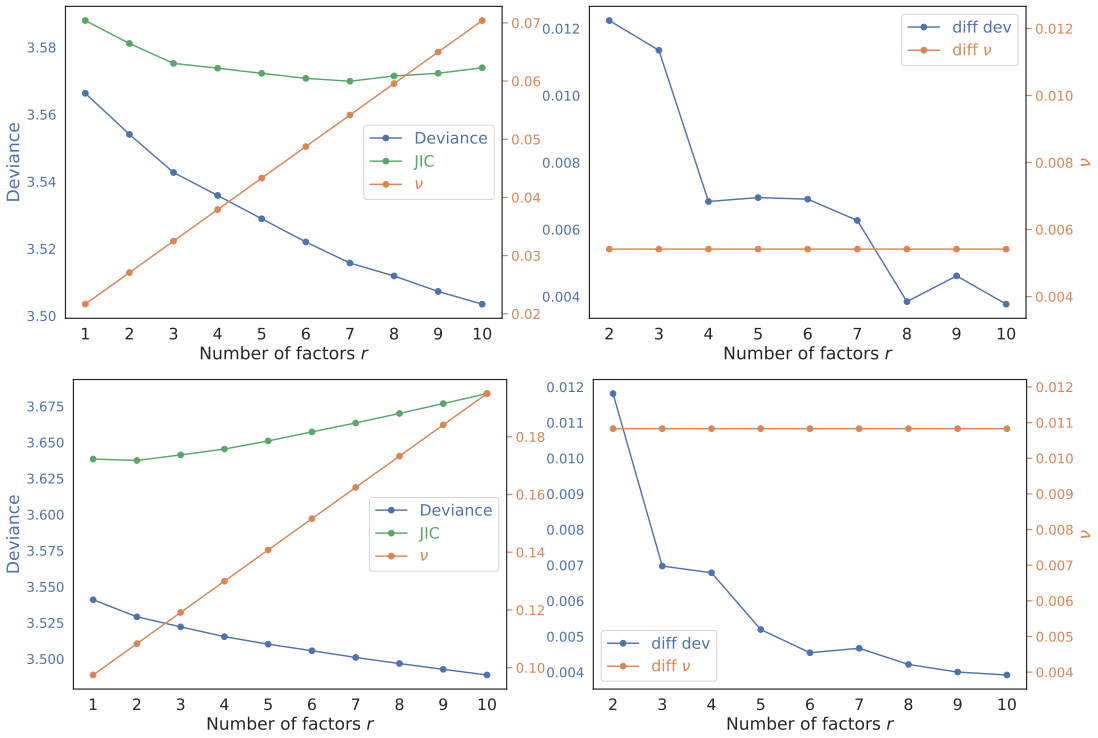

Figure A.75: The first and second rows show the results for GCATE-subset and GCATE-full, respectively. The right panel shows the deviance and the complexity penalty ν at different numbers of factors r, computed on the T4 cell type of the Lupus dataset. The JIC is computed with $c_{\text{JIC}} = 0.25$ and 0.5, respectively. The right panel shows the decrement of the deviance and the complexity at different numbers of factors r.

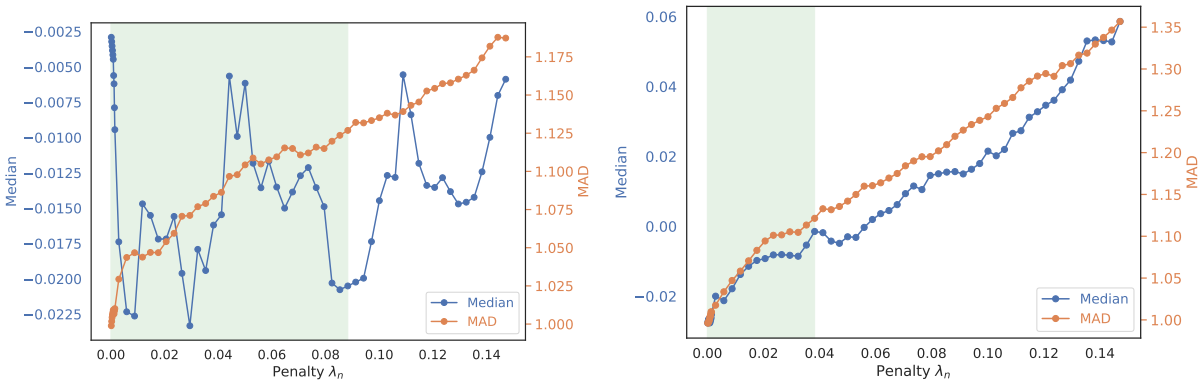

Figure A.76: The median and MAD of the z-statistics as a function of the regularization parameter λ_n computed from the T4 cell type of the Lupus dataset for GCATE-subset and GCATE-full analyses, respectively. The shaded region indicates feasible values of λ_n, for which the MADs of the corresponding test statistics are less than 1.13.

Results on all cell types

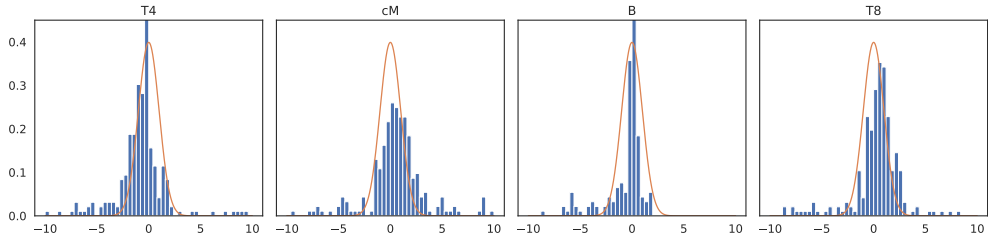

Figure A.77: Histograms of lupus z-statistics of CATE on T4, cM, B, and T8 cell types, when restricted to the top 250 highly variable genes. The preprocessing procedure is as described in Section 2.6, but with genes expressed less than 5 subjects excluded. The result on the NK cell type is not included because the fitting of CATE fails due to sparsity of the gene expressions.

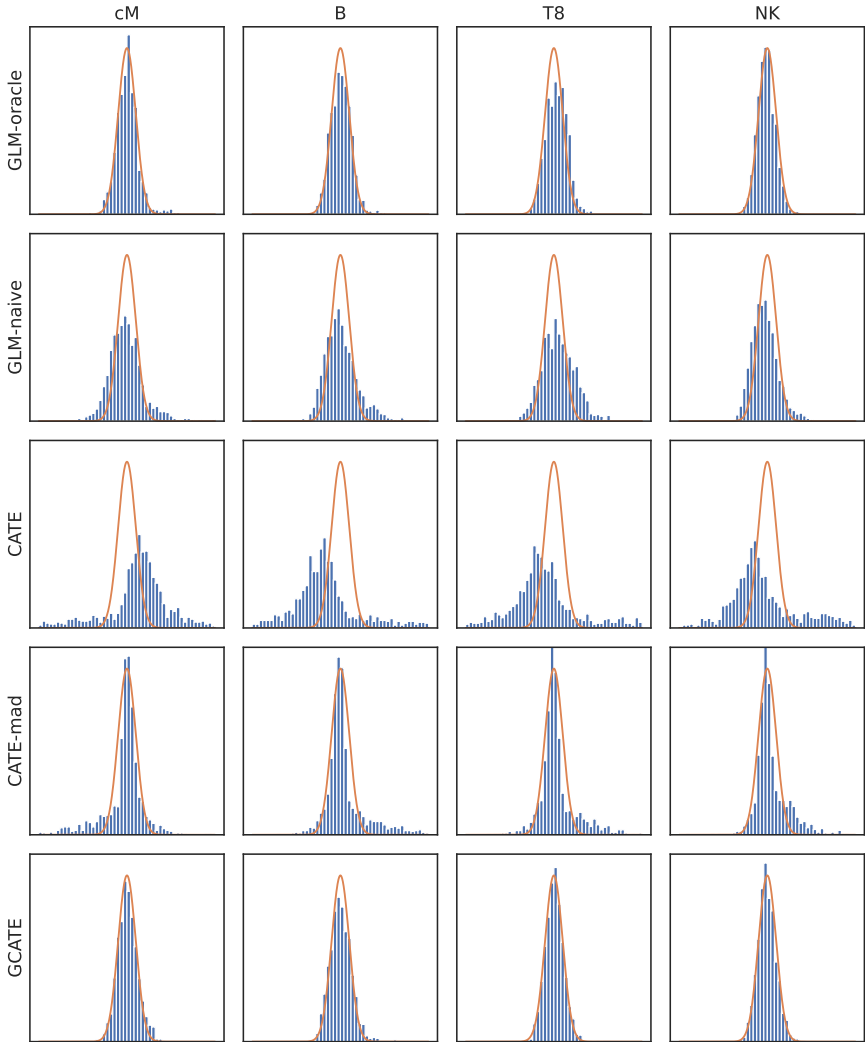

Figure A.78: Histograms of lupus z-statistics of different methods on cM, B, T8 and NK cell types.

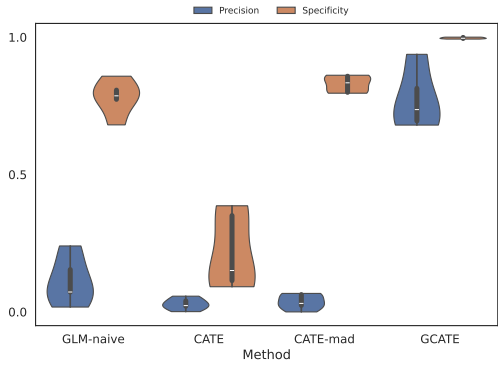

Figure A.79: The precision and specificity for four methods computed across 5 major cell types on the lupus datasets.

Figure A.710: The treemap plot produced by `rrvgo` [151] of GO enrichment analysis results on **(a)** significant genes by the GLM-oracle method; **(b)** significant genes by both the GLM-oracle and GCATE methods; and **(c)** significant genes by the CATE-mad method but not the GLM-oracle method.

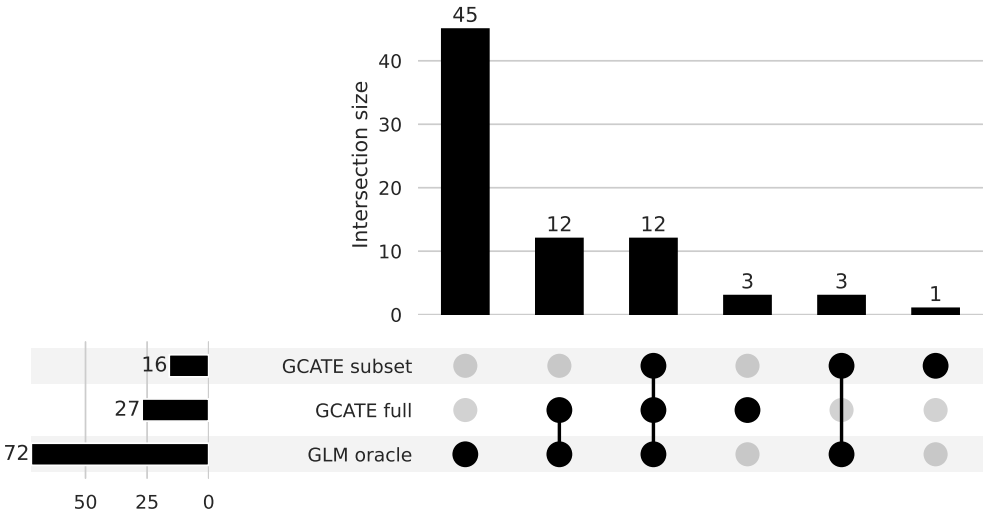

Figure A.711: Upset plot of the number of discoveries of GCATE (subset), GCATE (full) and GLM-oracle, with q-value cutoff 0.2. Here, "subset" and "full" indicate whether all of the measured covariates are used by the corresponding methods.

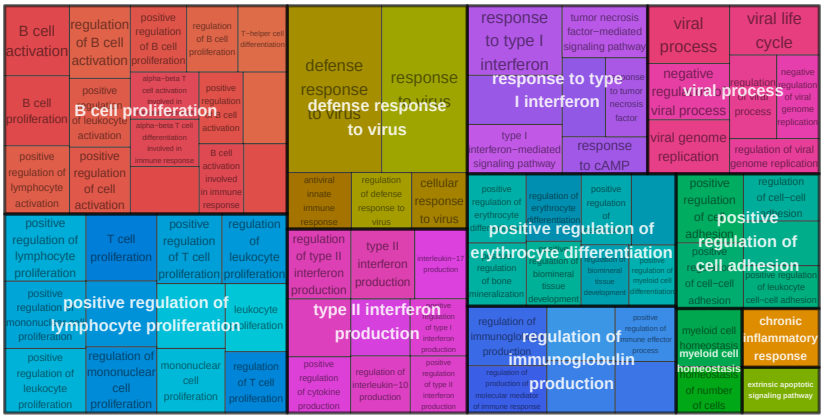

Figure A.712: The treemap plot produced by `rrvgo` [151] of GO enrichment analysis results on 24 significant genes by both the GLM and GCATE methods with all covariates included.

Appendix B

Causal Inference for Genomic Data with Multiple Heterogeneous Outcomes

B.1 Related work

In the context of causal inference, assessing causal effects on multiple outcomes requires accounting for the association among outcomes [148]. Earlier work on this problem relies on outcome modeling approaches based on linear mixed models or latent variable models [48, 113, 162, 163] and scaled linear models [106, 146]. Pocock et al. [138], Yoon et al. [180] study hypothesis testing of the treatment effects with the adjustment for multiplicities. Mattei et al. [117], Mealli and Pacini [121], Mealli et al. [122], Mercatanti et al. [123] use multiple outcomes, coupled with conditional independence assumptions, to address identification problems in causal studies with intermediate/post-treatment variables. In most of the aforementioned work, the number of outcomes is typically assumed to be low-dimensional.

The focus on multiple outcomes also shifts from outcome modeling to more general setups. Flanders and Klein [52] propose a general definition of causal effects, showing how it can be applied in the presence of multivariate outcomes for specific sub-populations of units or vectors of causal effects. For randomized experiments, Li and Ding [102] establish finite population central limit theorems in completely randomized experiments where the response variable may be multivariate and causal estimands of interest are defined as linear combinations of the potential outcomes. For observational studies with derived outcomes, Recently, Qiu et al. [139] propose an inverse probability weighting estimator for testing multiple average treatment effects (ATEs), which relies on the correct specification of the propensity score model and fast convergence rate of the propensity score estimation.

Although ATE is the most fundamental and popular causal estimand [70, 167], other estimands could be more robust to quantify the treatment effect between the counterfactual distributions. In the canonical setting with a single outcome, Athey et al. [6] explore the application of quantile methodologies to estimate the overall treatments within the context of randomized trials with heavy-tailed outcome distributions. For studies based on observational data, Belloni et al. [14] and Kallus et al. [81] introduce localized debiased machine learning techniques for quantile treatment effects (QTEs), which incorporates multiple sample partitioning. Concurrently, Chakrabortty et al. [23] study the estimation of QTEs within a semi-supervised framework. For a comprehensive

overview of QTE estimation literature, the reader is directed to the references contained therein. The above methods require sample splitting and/or metric entropy conditions to validate the asymptotic normality of the proposed estimators, even for a single outcome.

For the analysis of multiple outcomes extending beyond ATEs, Kennedy et al. [85] propose DR estimators designed to evaluate both scaled average treatment effects and scaled quantile effects. These estimands are used to rigorously test the global hypothesis that all treatment effects are equal. However, the asymptotic properties of the quantile-based estimators are not analyzed. Further, they only consider a low-dimensional set of outcomes. In contrast, the current paper focuses on high-dimensional settings when the number of outcomes could be potentially exponentially larger than the sample size. This also requires correctly addressing the issue of multiple hypothesis testing.

B.2 Proof in Section 3.2

B.2.1 Proof of Lemma 8

Proof of Lemma 8. Denote the scaled empirical process $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ by $\mathbb{G}_n$. By Lemma B.2.1, it follows that

$$\mathbb{E}\left[\max_{j=1,\dots,p} |\mathbb{G}_n g_j| \,\middle|\, g_1, \dots g_p\right] \lesssim \sqrt{\log p} \max_{1\le j\le p} \|g_j\|_{L_2} + \frac{(\log p)^{1-1/q}}{n^{1/2-1/q}} \|G\|_{L_q}.$$

Dividing $\sqrt{n}$ on both sides finishes the proof. □

Remark 13. We note that Lemma 8 also provides a probabilistic bound:

$$\max_{j=1,\dots,p} |(\mathbb{P}_n - \mathbb{P})g_j| = \mathcal{O}_{\mathbb{P}}\left(\left(\frac{\log p}{n}\right)^{1/2} \max_{1\le j\le p} \|g_j\|_{L_2} + \left(\frac{\log p}{n}\right)^{1-1/q} \|G\|_{L_q}\right),$$

by applying Markov inequality on the non-negative random variable $\max_{j=1,\dots,p} |(\mathbb{P}_n - \mathbb{P})g_j|$.

Remark 14 (Donsker condition)**.** Without an independent sample, similar bounds on the empirical process term can still be derived, provided certain complexity measures of the function class $\mathcal{F}_j$ that g_j belongs to are properly bounded. The complexity measure is related to the covering number of $\mathcal{F}_j$ under the L_2 norm induced by distribution Q, denoted by $N(\varepsilon, \mathcal{F}_j, L_2(Q))$. In particular, if for some envelope function F,

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}_j, L_2(Q))}\, \mathrm{d}\varepsilon < \infty,$$

and the supreme is taken over all probability distributions Q on $\mathcal{Z}$, then $\mathcal{F}_j$ is P-Donsker for every probability measure P such that $P^*F^2 < \infty$ under certain measurability conditions; see van der Vaart and Wellner [170, Theorem 2.5.2] for the exact conditions. To control single empirical process terms, a sufficient condition is that each $\mathcal{F}_j$ has a polynomial covering number:

$$\sup_Q N(\varepsilon, \mathcal{F}_j, L_2(Q)) \le \left(\frac{K}{\varepsilon}\right)^V, \qquad \forall\ 0 < \epsilon < 1,$$

where $K, V > 0$ are constants, which is similar to Chakrabortty et al. [23, Assumption 3.4]. If further the functions $g_j = \varphi_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \varphi_j(\boldsymbol{Z}; \mathbb{P})$ are uniformly bounded, by using a Bernstein-type

tail bound on the empirical process (e.g., van der Vaart and Wellner [170, Theorem 2.14.9]) and a union bound argument, we can obtain a similar upper bound on $\max_{j=1,\dots,p} |(\mathbb{P}_n - \mathbb{P})g_j|$ as in Lemma 1, with $\max_{1\le j\le p} \|g_j\|_{L_2}$ replaced by $\max_{1\le j\le p} \sup_{g\in\mathcal{F}_j} \|g\|_{L_2}$. When the function classes $\mathcal{F}_j = \mathcal{F}$ for all j are the same, this can be further improved to $\sup_{g\in\mathcal{F}} \|g\|_{L_2}$. However, verifying such assumptions on metric entropy in practice is challenging, and training the nuisance functions on an independent sample avoids this issue. Hence, in this paper, we use sample splitting instead of Donsker-type conditions.

If the Donsker class condition is assumed, Theorems 12 and 16 can be established without an independent sample. More specifically, the bias term $T_{\mathrm{R},j}$ in the decomposition (3.2.2) can be bounded under the same rate conditions on the product of two nuisance estimations in Theorems 12 and 16. This does not rely on independent sample splitting. Sample splitting is mainly used to control the empirical process term $T_{\mathrm{E},j}$ via Lemma 8 for the proof of Theorems 12 and 16, as described in Section 2.

B.2.2 Proof of Lemma 9

Proof of Lemma 9. Below, we condition on the event when conditions (1) and (2) hold. We begin by decomposing the error into two terms:

$$\begin{aligned}
\widehat{\sigma}_j^2 - \sigma_j^2 &= \mathbb{V}_n(\widehat{\varphi}_j) - \mathbb{V}[\varphi_j] \\
&= \mathbb{P}_n[(\widehat{\varphi}_j - \mathbb{P}_n[\widehat{\varphi}_j])^2] - \mathbb{P}[(\varphi_j - \mathbb{P}[\varphi_j])^2] \\
&= \mathbb{P}_n[\widehat{\varphi}_j^2] - (\mathbb{P}_n[\widehat{\varphi}_j])^2 - \mathbb{P}[\varphi_j^2] + (\mathbb{P}[\varphi_j])^2 \\
&= \mathbb{P}_n[\widehat{\varphi}_j^2 - \varphi_j^2] + (\mathbb{P}_n - \mathbb{P})[\varphi_j^2] - (\mathbb{P}_n[\widehat{\varphi}_j] - \mathbb{P}[\varphi_j])\,(\mathbb{P}_n[\widehat{\varphi}_j] + \mathbb{P}[\varphi_j]) \\
&= T_{1j} + T_{2j} + T_{3j}.
\end{aligned}$$

For the first term, we have

$$\max_{1\le j\le p} |T_{1j}| = \max_j |\mathbb{P}_n[(\widehat{\varphi}_j - \varphi_j)(\widehat{\varphi}_j + \varphi_j)]| \lesssim \max_j |\mathbb{P}_n[|\widehat{\varphi}_j - \varphi_j|]|\,,$$

where we use the boundedness of $(\widehat{\varphi}_j + \varphi_j)$'s envelope.

For the third term, we have

$$\max_{1\le j\le p} |T_{3j}| \lesssim \max_j |(\mathbb{P}_n - \mathbb{P})[\varphi_j]| + \max_j |(\mathbb{P}_n - \mathbb{P})[\widehat{\varphi}_j - \varphi_j]| + \max_j |\mathbb{P}[\widehat{\varphi}_j - \widehat{\varphi}_j]|.$$

Combining the above results yields that

$$\begin{aligned}
\max_{1\le j\le p} |\widehat{\sigma}_j^2 - \sigma_j^2| \lesssim \max_j |(\mathbb{P}_n - \mathbb{P})[\widehat{\varphi}_j - \varphi_j]| + \max_j |(\mathbb{P}_n - \mathbb{P})[|\widehat{\varphi}_j - \varphi_j|]| \\
+ \max_{k=1,2} \max_j \left|(\mathbb{P}_n - \mathbb{P})[\varphi_j^k]\right| + \max_j \mathbb{P}[|\widehat{\varphi}_j - \varphi_j|],
\end{aligned}$$

with probability tending to one.

Define $\Psi = \max_{1\le j\le p} |\widehat{\varphi}_j - \varphi_j|$ and $\Phi = \max_{1\le j\le p} |\varphi_j|$. By Lemma 8, it follows that

$$\begin{aligned}
\max_{1\le j\le p} |\widehat{\sigma}_j^2 - \sigma_j^2| \lesssim & \left(\frac{\log p}{n}\right)^{1/2} \max_j \|\widehat{\varphi}_j - \varphi_j\|_{L_2} + \left(\frac{\log p}{n}\right)^{1-1/q} \|\Psi\|_{L_q} \\
& + \left(\frac{\log p}{n}\right)^{1/2} \max_{k=1,2} \max_j \|\varphi_j^k\|_{L_2} + \left(\frac{\log p}{n}\right)^{1-1/q} \max_{k=1,2} \|\Phi^k\|_{L_q} \\
& + \max_j \|\widehat{\varphi}_j - \varphi_j\|_{L_1} \\
= & \left(\frac{\log p}{n}\right)^{1/2} \max_j \left(\|\widehat{\varphi}_j - \varphi_j\|_{L_2} + \max_{k=1,2} \|\varphi_j^k\|_{L_2} \right) + \max_j \|\widehat{\varphi}_j - \varphi_j\|_{L_1} \\
& + \left(\frac{\log p}{n}\right)^{1-1/q} \left(\|\Psi\|_{L_q} + \max_{k=1,2} \|\Phi^k\|_{L_q} \right),
\end{aligned}$$

which holds with probability at least $1 - n^{-c}$. □

B.2.3 Helper lemmas

Lemma B.2.1 (Maximal inequality adapted from Proposition B.1 of Kuchibhotla and Patra [90])**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be mean zero independent random variables in $\mathbb{R}^p$ for $p \ge 1$. Suppose there exists $q \in \mathbb{N}$ such that for all $i \in \{1, \ldots, n\}$

$$\mathbb{E}\left[\xi_i^q\right] < \infty \quad \text{where} \quad \xi_i := \max_{1\le j\le p} |X_{i,j}|$$

and $\boldsymbol{X}_i := (X_{i,1}, \ldots, X_{i,p})^\top$. If $V_{n,p} := \max_{1\le j\le p} \sum_{i=1}^n \mathbb{E}\left[X_{i,j}^2\right]$, then

$$\mathbb{E}\left[\max_{1\le j\le p} \left|\sum_{i=1}^n X_{i,j}\right|\right] \le \sqrt{6V_{n,p}\log(1+p)} + \sqrt{2}(3\log(1+p))^{1-1/q} \left(2\sum_{i=1}^n \mathbb{E}\left[\xi_i^q\right]\right)^{1/q}.$$

B.3 Identification conditions

B.3.1 Proof of Proposition 10

Proof of Proposition 10. Note that

$$\begin{aligned}
& \mathbb{E}(\widetilde{\boldsymbol{Y}} \mid A = a, \boldsymbol{W}) \\
=& \mathbb{E}(\widetilde{\boldsymbol{Y}}(a) \mid A = a, \boldsymbol{W}) \\
=& \mathbb{E}(\widetilde{\boldsymbol{Y}}(a) \mid \boldsymbol{W}) \\
=& \mathbb{E}[\mathbb{E}[\widetilde{\boldsymbol{Y}}(a) \mid \boldsymbol{W}, \boldsymbol{S}(a)] \mid \boldsymbol{W}] \\
=& \mathbb{E}[\boldsymbol{Y}(a) \mid \boldsymbol{W}] + \mathbb{E}[\boldsymbol{\Delta}_m(a) \mid \boldsymbol{W}].
\end{aligned}$$

where the first equality is from the consistency and positivity assumption, and the second equality is from the unmeasured confounder assumption. From the asymptotic unbiasedness as in Definition 1, it follows that

$$\mathbb{E}[\mathbb{E}[\widetilde{\boldsymbol{Y}} \mid A = a, \boldsymbol{W}]] = \mathbb{E}[\boldsymbol{Y}(a)] + \mathbb{E}[\boldsymbol{\Delta}_m(a)] = \mathbb{E}[\boldsymbol{Y}(a)] + \boldsymbol{o}(1), \tag{B.3.1}$$

where the little-o notation is with respect to m and uniform in p. This completes the proof. □

B.3.2 Proof of Lemma 15

Proof of Lemma 15. Define $\widetilde{M}_j(\theta) = \mathbb{E}[F_j(\widetilde{Y}_j(a), \theta)]$. From the assumption, we have that

$$\max_{j\in[p]} |M_j(\theta) - \widetilde{M}_j(\theta)| = \max_{j\in[p]} |\mathbb{E}[\Delta_{mj}(a,\theta)]|.$$

By Taylor expansion, we have

$$M_j(\widetilde{\theta}_{aj}) - M_j(\theta_{aj}) = M_j'(\bar{\theta}_{aj})(\widetilde{\theta}_{aj} - \theta_{aj}).$$

for some $\bar{\theta}_{aj}$ between $\widetilde{\theta}_{aj}$ and θ_{aj}. These two results imply that

$$\begin{aligned}
\max_{j\in[p]} |\widetilde{\theta}_{aj} - \theta_{aj}| &= \max_{j\in[p]} \frac{1}{|M_j'(\bar{\theta}_{aj})|} |M_j(\widetilde{\theta}_{aj}) - M_j(\theta_{aj})| \\
&\le \frac{1}{c} \max_{j\in[p]} |M_j(\widetilde{\theta}_{aj})| \\
&= \frac{1}{c} \max_{j\in[p]} |\widetilde{M}_j(\widetilde{\theta}_{aj}) + M_j(\widetilde{\theta}_{aj}) - \widetilde{M}_j(\widetilde{\theta}_{aj})| \\
&= \frac{1}{c} \max_{j\in[p]} |M_j(\widetilde{\theta}_{aj}) - \widetilde{M}_j(\widetilde{\theta}_{aj})|
\end{aligned}$$

Similar to the proof of Proposition 10, we have that

$$\widetilde{M}_j(\theta) = \mathbb{E}[\mathbb{E}[F_j(\widetilde{Y}_j, \theta) \mid A = a, \boldsymbol{W}]] = M_j(\theta) + \mathbb{E}[\Delta_{mj}(a,\theta)].$$

Thus we have

$$\max_{j\in[p]} |\widetilde{\theta}_{aj} - \theta_{aj}| \le \frac{1}{c} \max_{j\in[p]} |\mathbb{E}[\Delta_{mj}(a, \widetilde{\theta}_{aj})]| \le \frac{1}{c} \max_{j\in[p]} \sup_{\theta\in\mathcal{B}(\theta_{aj},\delta)} |\mathbb{E}[\Delta_{mj}(a,\theta)]| \lesssim \delta_m \to 0$$

as $m \to \infty$. □

B.4 Doubly robust estimation

B.4.1 Proof of Lemma 11

Proof of Lemma 11. For all $i = 1, 2$, and $j = 1, \ldots, p$, define

$$\psi_{aij} = \mathbb{E}[Y_j(a)^i], \qquad \widetilde{\psi}_{aij} = \mathbb{E}[\mathbb{E}[\widetilde{Y}_j^i \mid A = a, \boldsymbol{W}]].$$

From Proposition 10 we have that $\psi_{aij} = \widetilde{\psi}_{aij} + o(1)$ as $m \to \infty$. It follows that

$$\begin{aligned}
\widetilde{\tau}_j &= \frac{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=1, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=0, \boldsymbol{W})]}{\sqrt{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j^2 \mid A=0, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=0, \boldsymbol{W})]^2}} \\
&= \frac{\widetilde{\psi}_{11j} - \widetilde{\psi}_{01j}}{\sqrt{\widetilde{\psi}_{02j} - \widetilde{\psi}_{01j}^2}} \\
&= \frac{\psi_{11j} - \psi_{01j} + o(1)}{\sqrt{\psi_{02j} - \psi_{01j}^2 + o(1)}} \\
&= \frac{\psi_{11j} - \psi_{01j}}{\sqrt{\psi_{02j} - \psi_{01j}^2}} + o(1) \\
&= \tau_j + o(1),
\end{aligned}$$

where the second last equality holds because $\mathbb{V}[Y_j(0)] = \psi_{02j} - \psi_{01j}^2 > 0$ as assumed. □

B.4.2 Proof of Theorem 12

Proof of Theorem 12. In this proof we will abbreviate τ_j^{STE} as τ_j and denote

$$\widetilde{\tau}_j = \frac{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=1, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=0, \boldsymbol{W})]}{\sqrt{\mathbb{E}[\mathbb{E}(\widetilde{Y}_j^2 \mid A=0, \boldsymbol{W})] - \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=0, \boldsymbol{W})]^2}}.$$

We split the proof into three parts, conditioned on the event when the assumptions hold (which holds with probability tending to one).

Part (1) Individual ATE estimators. Let $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \beta_{2j})^\top$ with

$$\beta_{0j} = \mathbb{E}[Y_j(0)],\ \beta_{1j} = \mathbb{E}[Y_j(1)],\ \beta_{2j} = \mathbb{E}[Y_j(0)^2],$$

Let $\widetilde{\boldsymbol{\beta}}_j = \left(\widetilde{\beta}_{0j}, \widetilde{\beta}_{1j}, \widetilde{\beta}_{2j}\right)^\top$ with

$$\widetilde{\beta}_{0j} = \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=0, \boldsymbol{W})],\ \widetilde{\beta}_{1j} = \mathbb{E}[\mathbb{E}(\widetilde{Y}_j \mid A=1, \boldsymbol{W})],\ \widetilde{\beta}_{2j} = \mathbb{E}[\mathbb{E}(\widetilde{Y}_j^2 \mid A=0, \boldsymbol{W})],$$

and define the corresponding estimator $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{0j}, \widehat{\beta}_{1j}, \widehat{\beta}_{2j})^\top$ for

$$\widehat{\beta}_{0j} = \mathbb{P}_n\left\{\widetilde{\phi}_{01j}(\boldsymbol{Z}; \widehat{\pi}, \widehat{\boldsymbol{\mu}})\right\}, \qquad \widehat{\beta}_{1j} = \mathbb{P}_n\left\{\widetilde{\phi}_{11j}(\boldsymbol{Z}; \widehat{\pi}, \widehat{\boldsymbol{\mu}})\right\}, \qquad \widehat{\beta}_{2j} = \mathbb{P}_n\left\{\widetilde{\phi}_{02j}(\boldsymbol{Z}; \widehat{\pi}, \widehat{\boldsymbol{\mu}})\right\}.$$

Let $\boldsymbol{\phi}_j = (\widetilde{\phi}_{01j}, \widetilde{\phi}_{11j}, \widetilde{\phi}_{02j})^\top$ and

$$\delta_{m0} = \max_{1 \le j \le p} |\mathbb{E}[\Delta_{m1j}(0)]|, \qquad \delta_{m1} = \max_{1 \le j \le p} |\mathbb{E}[\Delta_{m1j}(1)]|, \qquad \delta_{m2} = \max_{1 \le j \le p} |\mathbb{E}[\Delta_{m2j}(0)]|.$$

Similar to Lemma B.4.1, we can show that

$$\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j = (\mathbb{P}_n - \mathbb{P})\left[\boldsymbol{\phi}_j(\boldsymbol{Z}; \pi, \boldsymbol{\mu})\right] + \boldsymbol{\epsilon}_j$$

where $\max_{1 \le j \le p} \|\boldsymbol{\epsilon}_j\|_\infty = \mathcal{O}(r_{np})$ and

$$r_{np} = (\log p)^{1/2} n^{-(1/2 + \alpha \wedge \beta)} + \log p / n + n^{-(\alpha+\beta)}.$$

Part (2) Bounding $\widehat{\tau}_j - \widetilde{\tau}_j$. Now since $\widehat{\tau}_j = h(\widehat{\boldsymbol{\beta}}_j) := (\widehat{\beta}_1 - \widehat{\beta}_0)(\widehat{\beta}_2 - \widehat{\beta}_0^2)^{-1/2}$, an application of Taylor's expansion yields

$$\begin{aligned}
&\widehat{\tau}_j - \widetilde{\tau}_j \\
&= (\mathbb{P}_n - \mathbb{P}) \left[\nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\phi}_j(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right] + \nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\epsilon}_j + \mathcal{O}_{\mathbb{P}}(\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2^2) \\
&= \mathbb{P}_n \left[\frac{\widetilde{\phi}_{11j} - \widetilde{\phi}_{01j}}{\sqrt{\widetilde{\beta}_{2j} - \widetilde{\beta}_{0j}^2}} - \widetilde{\tau}_j \left\{ \frac{\widetilde{\phi}_{02j} + \widetilde{\beta}_{2j} - 2\widetilde{\beta}_{0j}\widetilde{\phi}_{01j}}{2\left(\widetilde{\beta}_{2j} - \widetilde{\beta}_{0j}^2\right)} \right\} \right] + \nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\epsilon}_j + \mathcal{O}_{\mathbb{P}}(\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2^2). \quad \text{(B.4.1)}
\end{aligned}$$

We denote the first term by $\mathbb{P}_n\{\widetilde{\varphi}_j\}$. For remainder term $\nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\epsilon}_j$, by proof of Proposition 10 we have

$$\max_{1 \le j \le p} \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2 \le \sqrt{3} \max_{1 \le j \le p} \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_\infty = \sqrt{3}\delta_m$$

Thus, from the bounded variance assumption that $\mathbb{V}[Y_j(0)] = \beta_{j2} - \beta_{j0}^2 \ge c$ and

$$\max_{1 \le j \le p} |\mathbb{E}[\Delta_{mkj}(a)]| = \delta_m \to 0,$$

when m is large, $\max_{1 \le j \le p} \|\nabla h(\widetilde{\boldsymbol{\beta}}_j)\|_2 \le C$ for some constant C. We have

$$\begin{aligned}
|\nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\epsilon}_j| &\le \|\nabla h(\widetilde{\boldsymbol{\beta}}_j)\|_2 \|\boldsymbol{\epsilon}_j\|_2 \le \sqrt{3} C \|\boldsymbol{\epsilon}_j\|_\infty, \\
\max_{1 \le j \le p} |\nabla h(\widetilde{\boldsymbol{\beta}}_j)^\top \boldsymbol{\epsilon}_j| &\lesssim \mathbb{E}\left[\max_{1 \le j \le p} \|\boldsymbol{\epsilon}_j\|_\infty \right] = \mathcal{O}(r_{np}).
\end{aligned}$$

The second-order remainder is bounded as

$$\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2^2 \lesssim \|(\mathbb{P}_n - \mathbb{P}) \left[\boldsymbol{\phi}_j(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]\|_2^2 + \|\boldsymbol{\epsilon}_j\|_\infty^2,$$

where we have

$$\begin{aligned}
&\|(\mathbb{P}_n - \mathbb{P}) \left[\boldsymbol{\phi}_j(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]\|_2^2 \\
&= |(\mathbb{P}_n - \mathbb{P}) \left[\widetilde{\phi}_{01j}(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]|^2 + |(\mathbb{P}_n - \mathbb{P}) \left[\widetilde{\phi}_{11j}(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]|^2 + |(\mathbb{P}_n - \mathbb{P}) \left[\widetilde{\phi}_{02j}(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]|^2
\end{aligned}$$

Since $\widetilde{\phi}_{01j}, \widetilde{\phi}_{11j}, \widetilde{\phi}_{02j}$ are all bounded, by Lemma 8 with $q = \infty$ we have

$$\max_{1 \le j \le p} |(\mathbb{P}_n - \mathbb{P}) \left[\widetilde{\phi}_{01j}(\boldsymbol{Z}; \pi, \boldsymbol{\mu}) \right]|^2 = \mathcal{O}\left(\frac{\log p}{n} \right)$$

and same bound holds for $\widetilde{\phi}_{11j}, \widetilde{\phi}_{02j}$ as well. This together with $\mathbb{E}[\max_{1 \le j \le p} \|\boldsymbol{\epsilon}_j\|_\infty] = \mathcal{O}(r_{np})$ implies (note that r_{np} includes the $\log p/n$ term)

$$\max_{1 \le j \le p} \|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2^2 = \mathcal{O}\left(\frac{\log p}{n} + r_{np}^2 \right) = \mathcal{O}(r_{np}). \quad \text{(B.4.2)}$$

Combining the above bounds with (B.4.1) implies that

$$\widehat{\tau}_j - \widetilde{\tau}_j = \mathbb{P}_n\{\widetilde{\varphi}_j\} + \varepsilon_j' \quad \text{(B.4.3)}$$

where the residual terms satisfy $\max_{j \in [p]} |\varepsilon_j'| = \mathcal{O}((\log p)^{1/2} n^{-(1/2 + \alpha \wedge \beta)} + \log p/n + n^{-(\alpha+\beta)})$.

Part (3) Bounding $\widetilde{\tau}_j - \tau_j$. By Taylor expansion, we also have

$$\widetilde{\tau}_j - \tau_j = h(\widetilde{\boldsymbol{\beta}}_j) - h(\boldsymbol{\beta}_j) = \nabla h(\bar{\boldsymbol{\beta}}_j)^\top (\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$$

where $\bar{\boldsymbol{\beta}}_j$ lies between $\widetilde{\boldsymbol{\beta}}_j$ and $h(\boldsymbol{\beta}_j)$ so it also lies in the ball $\mathcal{B}_j(\boldsymbol{\beta}_j, \delta_m)$ where $\delta_m = \max\{\delta_{m0}, \delta_{m1}, \delta_{m2}\}$ is the maximum bias of derived outcomes. When m is large, we also have $\|\nabla h(\bar{\boldsymbol{\beta}}_j)\|_2 \leq C$. This implies

$$\max_{1\leq j\leq p} |\widetilde{\tau}_j - \tau_j| \lesssim \|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2 \lesssim \delta_m.$$

Combining the above bounds with (B.4.3), we have

$$\widehat{\tau}_j - \tau_j = \mathbb{P}_n\{\widetilde{\varphi}_j\} + \varepsilon_j,$$

where the residual terms satisfy $\max_{j\in[p]} |\varepsilon_j| = \mathcal{O}((\log p)^{1/2} n^{-(1/2+\alpha\wedge\beta)} + \log p/n + n^{-(\alpha+\beta)} + \delta_m)$. □

B.4.3 Proof of Proposition 14

Proof of Proposition 14. We next verify that the conditions in Lemma 9 hold the event when the assumptions hold. Recall the centered influence function $\widetilde{\varphi}_j = \widetilde{\varphi}_j^{\mathrm{STE}}$ defined in Theorem 12. We write $\widetilde{\varphi}_j(\boldsymbol{Z}; \mathbb{P}) = \widetilde{\varphi}_j(\boldsymbol{Z}; \pi, \boldsymbol{\mu})$ and $\widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) = \widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\pi}, \widehat{\boldsymbol{\mu}})$. By the boundedness assumption, we have that

$$\max_{j\in[p]} |\widetilde{\varphi}_j(\boldsymbol{Z}; \mathbb{P}) + \widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}})| = \mathcal{O}(1).$$

From the proof of Lemma B.4.1, the individual influence functions satisfy that

$$\max_{1\leq j\leq p} \|\widetilde{\phi}_{akj}(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \widetilde{\phi}_{akj}(\boldsymbol{Z}; \mathbb{P})\|_{L_2} = \mathcal{O}(n^{-\alpha\wedge\beta}), \qquad a = 0, 1,\ k = 0, 1, 2,$$

The estimation error $\|\widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \widetilde{\varphi}_j(\boldsymbol{Z}; \mathbb{P})\|_2$ then depends on the slowest rate among $\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2, \widehat{\tau}_j - \widetilde{\tau}_j$ and $\widetilde{\phi}_{akj}(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \widetilde{\phi}_{akj}(\boldsymbol{Z}; \mathbb{P})$. By the proof in Appendix B.4.2, we have

$$\max_{1\leq j\leq p} \|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2 = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} + r_{np}\right),$$
$$\max_{1\leq j\leq p} |\widehat{\tau}_j - \widetilde{\tau}_j| = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} + r_{np}\right).$$

Combining this with the positivity assumption of $\mathbb{V}[\widetilde{Y}_j(0)]$ implies that

$$\max_{1\leq j\leq p} \|\widetilde{\varphi}_j(\boldsymbol{Z}; \widehat{\mathbb{P}}) - \widetilde{\varphi}_j(\boldsymbol{Z}; \mathbb{P})\|_{L_2} = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} + n^{-\alpha\wedge\beta}\right). \tag{B.4.4}$$

Therefore, by applying Lemma 9, we have that $\max_{j\in[p]} |\widehat{\sigma}_j^2 - \sigma_j^2| = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n} + n^{-\alpha\wedge\beta})$ as $m, n, p \to \infty$ such that $\log p = o(n^{\min(\frac{1}{2}, 2\alpha, 2\beta)})$. □

B.4.4 Proof of Theorem 16

Proof of Theorem 16. We condition on the event when the assumptions hold. In the proof of Proposition 15, we show

$$\max_{j\in[p]}|\widetilde{\theta}_{aj}-\theta_{aj}| \le \frac{1}{c}\max_{j\in[p]}\sup_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)}|\mathbb{E}[\Delta_{mj}(a,\theta)]| \lesssim \delta_m.$$

We first inspect the estimation error of counterfactual quantiles for $a\in\{0,1\}$. Note that Condition 2 implies that

$$\mathbb{P}(\cap_{j=1}^{p}\{\widehat{\theta}_{aj}^{\text{init}}\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)\})\to 1.$$

Also, Conditions 1 and 2 imply that

$$\max_{j\in[p]}\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}}) = \mathcal{O}(1) \tag{B.4.5}$$

$$\widehat{L}_{np} := \max_{j\in[p]}|\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1} - f_{aj}(\widetilde{\theta}_{aj})^{-1}| = \mathcal{O}(n^{-\kappa}) \tag{B.4.6}$$

We begin by writing the estimation error as

$$\begin{aligned}
\widehat{\theta}_{aj}-\widetilde{\theta}_{aj} &= \widehat{\theta}_{aj}^{\text{init}} + \frac{1}{\widehat{f}_j(\widehat{\theta}_{aj}^{\text{init}})}\mathbb{P}_n[\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})] - \widetilde{\theta}_{aj} \\
&= f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})] \\
&\quad + \widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1}(\mathbb{P}_n-\mathbb{P})[\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})] \\
&\quad + \widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1}\mathbb{P}[\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})] \\
&\quad + (\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj}) + \widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})] - f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})] \\
&=: f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})] + T_{j1} + T_{j2} + T_{j3}.
\end{aligned}$$

Next, we analyze the three residual terms separately.

Part (1) T_{j1}. This empirical process term can be uniformly bounded using Lemma B.4.2 with (B.4.5):

$$\begin{aligned}
\max_{j\in[p]}|T_{j1}| &= \mathcal{O}(\max_{j\in[p]}|(\mathbb{P}_n-\mathbb{P})\{\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})\}|) \\
&= \mathcal{O}\left(\sqrt{\frac{\log p}{n}}n^{-\alpha\wedge\beta}+\frac{\log p}{n}\right).
\end{aligned}$$

Part (2) T_{j2}. This bias term can be uniformly bounded using Lemma B.4.2 with (B.4.5):

$$\begin{aligned}
\max_{j\in[p]}|T_{j2}| &= \mathcal{O}(\max_{j\in[p]}|\mathbb{P}\{\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})\}|) \\
&= \mathcal{O}\left(n^{-(\alpha+\beta)}\right).
\end{aligned}$$

Part (3) T_{j3}**.** This extra bias term can be decomposed into two terms:

$$\begin{aligned}
T_{j3} &= (\widehat{\theta}_{aj}^{\text{init}} - \widetilde{\theta}_{aj}) + \widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})] - f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] \\
&= (\widehat{\theta}_{aj}^{\text{init}} - \widetilde{\theta}_{aj}) + f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] \\
&\quad + f_{aj}(\widetilde{\theta}_{aj})^{-1}(\mathbb{P}_n - \mathbb{P})[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] \\
&\quad + [\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\text{init}})^{-1} - f_{aj}(\widetilde{\theta}_{aj})^{-1}]\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})],
\end{aligned}$$

where we use the fact that $\mathbb{P}\{\omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})\} = 0$. From Lemma B.4.3 and (B.4.6), we have that

$$\begin{aligned}
\max_{j\in[p]} |T_{j3}| &= \mathcal{O}(n^{-2\gamma}) + \mathcal{O}\left(\sqrt{\frac{\log p}{n}} n^{-\gamma/2} + \frac{\log p}{n}\right) \\
&\quad + \mathcal{O}\left(n^{-\kappa}\left(\frac{\log p}{n} + \sqrt{\frac{\log p}{n}} + n^{-\gamma}\right)\right) \\
&= \mathcal{O}\left(n^{-2\gamma} + n^{-(\gamma+\kappa)} + \frac{\log p}{n} + \sqrt{\frac{\log p}{n}} n^{-\frac{\gamma}{2}\wedge\kappa}\right).
\end{aligned}$$

Combining the three terms yields

$$\widehat{\theta}_{aj} - \widetilde{\theta}_{aj} = f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] + \mathcal{O}\left(\frac{(\log p)^{1/2}}{n^{1/2+\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2}}} + \frac{\log p}{n} + n^{-(\alpha+\beta)\wedge(\gamma+\kappa)\wedge(2\gamma)}\right).$$

Setting $\widetilde{\varphi}_j^{\text{QTE}} = f_{1j}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}_n[\omega_{1j}(\boldsymbol{Z}, \widetilde{\theta}_{1j})] - f_{0j}(\widetilde{\theta}_{0j})^{-1}\mathbb{P}_n[\omega_{0j}(\boldsymbol{Z}, \widetilde{\theta}_{0j})]$ gives that

$$\widehat{\tau}_j^{\text{QTE}} - \widetilde{\tau}_j^{\text{QTE}} = \mathbb{P}_n\{\widetilde{\varphi}_j^{\text{QTE}}\} + \mathcal{O}\left(\frac{(\log p)^{1/2}}{n^{1/2+\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2}}} + \frac{\log p}{n} + n^{-(\alpha+\beta)\wedge(\gamma+\kappa)\wedge(2\gamma)}\right)$$

Recall we have

$$\max_{j\in[p]} |\widetilde{\theta}_{aj} - \theta_{aj}| \leq \frac{1}{c}\max_{j\in[p]} \sup_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)} |\mathbb{E}[\Delta_{mj}(a,\theta)]| \lesssim \delta_m,$$

which implies

$$\max_{j\in[p]} |\widetilde{\tau}_j^{\text{QTE}} - \tau_j^{\text{QTE}}| = \mathcal{O}(\delta_m).$$

Because the above results hold with probability tending to one, it further implies that

$$\widehat{\tau}_j^{\text{QTE}} - \tau_j^{\text{QTE}} = \mathbb{P}_n\{\widetilde{\varphi}_j^{\text{QTE}}\} + \mathcal{O}_{\mathbb{P}}\left(\frac{(\log p)^{1/2}}{n^{1/2+\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2}}} + \frac{\log p}{n} + n^{-(\alpha+\beta)\wedge(\gamma+\kappa)\wedge(2\gamma)} + \delta_m\right),$$

which finishes the proof. □

B.4.5 Proof of Proposition 17

Proof of Proposition 17. From Theorem 16, when $\log p = o(n^{2(\frac{1}{4}\wedge\alpha\wedge\beta\wedge\frac{\gamma}{2})})$ the asymptotic normality follows immediately from the triangular-array central limit theorem in Lemma B.4.4.

We next verify that the conditions in Lemma 9 hold under the assumptions. For simplicity, we drop the superscript QTE from $\widetilde{\varphi}_j^{\mathrm{QTE}}$ and write $\widetilde{\varphi}_j$. By the boundedness assumption, we have that

$$\max_{j\in[p]} |\widetilde{\varphi}_j(\boldsymbol{Z};\mathbb{P}) + \widetilde{\varphi}_j(\boldsymbol{Z};\widehat{\mathbb{P}})| = \mathcal{O}(1).$$

Notice that

$$\begin{aligned}
&\widetilde{\varphi}_j(\boldsymbol{Z};\widehat{\mathbb{P}}) - \widetilde{\varphi}_j(\boldsymbol{Z};\mathbb{P}) \\
&= \sum_{a\in\{0,1\}} \frac{1}{\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\mathrm{init}})}\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) - \frac{1}{f_{aj}(\widetilde{\theta}_{aj})}\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj}) \\
&= \sum_{a\in\{0,1\}} \left[\left(\frac{1}{\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\mathrm{init}})} - \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \right) \widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) + \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \left(\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) - \omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) \right) \right. \\
&\left. + \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \left(\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) - \omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj}) \right) \right].
\end{aligned}$$

Then we have that

$$\begin{aligned}
&\max_{j\in[p]} \mathbb{P}[|\widetilde{\varphi}_j(\boldsymbol{Z};\widehat{\mathbb{P}}) - \varphi_j(\boldsymbol{Z};\mathbb{P})|] \\
&\le \max_{j\in[p]} \|\widetilde{\varphi}_j(\boldsymbol{Z};\widehat{\mathbb{P}}) - \widetilde{\varphi}_j(\boldsymbol{Z};\mathbb{P})\|_{L_2} \\
&\le \max_{j\in[p]} \sum_{a\in\{0,1\}} \left[\left| \frac{1}{\widehat{f}_{aj}(\widehat{\theta}_{aj}^{\mathrm{init}})} - \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \right| \|\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}})\|_{L_2} + \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \|\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) - \omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}})\|_{L_2} \right. \\
&\left. + \frac{1}{f_{aj}(\widetilde{\theta}_{aj})} \|\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\mathrm{init}}) - \omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})\|_{L_2} \right] \\
&= \mathcal{O}(n^{-\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2}}).
\end{aligned}$$

where in the last equality, we use (B.4.11) and Lemma B.4.3. Therefore, by applying Lemma 9, we have that $\max_{j\in[p]} |\widehat{\sigma}_j^2 - \widetilde{\sigma}_j^2| = \mathcal{O}_{\mathbb{P}}(\log p/n + (\log p)^{1/2} n^{-(1/2+\alpha\wedge\beta\wedge\kappa\wedge\gamma/2)} + n^{-\alpha\wedge\beta\wedge\kappa\wedge\gamma/2})$ as $m, n, p \to \infty$ such that $\log p = o(n^{2(\frac{1}{4}\wedge\alpha\wedge\beta\wedge\kappa\wedge\frac{\gamma}{2})})$. □

B.4.6 Helper lemmas

Recall that the uncentered influence function is defined as $\phi_j(\boldsymbol{Z};\mathbb{P}) = \varphi_j(\boldsymbol{Z};\mathbb{P}) + \tau_j(\mathbb{P})$. The lemma below provides results for the asymptotic normality of the one-step estimator of $\mathbb{E}[Y_j(a)]$ in the setting of multiple derived outcomes. It can be analogously adapted to the one of ATE, which we omit for brevity.

Lemma B.4.1 (Counterfactual expectation)**.** For $a \in \{0,1\}$, suppose that Assumptions 5–7 hold and $\widetilde{\boldsymbol{Y}}(a)$ is asymptotically unbiased to $\boldsymbol{Y}(a)$. For $j = 1,\dots,p$, define $\tau_{aj} = \mathbb{E}[Y_j(a)]$, $\pi_a(\boldsymbol{W}) = \mathbb{P}(A = a \mid \boldsymbol{W})$ and $\mu_{aj}(\boldsymbol{W}) = \mathbb{E}[\widetilde{Y}_j \mid A = a, \boldsymbol{W}]$, the uncentered influence function as

$$\phi_{aj}(\boldsymbol{Z};\pi,\boldsymbol{\mu}) = \frac{\mathbb{1}\{A=a\}}{\pi_a(\boldsymbol{W})}(\widetilde{Y}_j - \mu_{aj}(\boldsymbol{W})) + \mu_{aj}(\boldsymbol{W}),$$

and the one-step estimator as $\widehat{\tau}_{aj} = \mathbb{P}_n[\phi_{aj}(\boldsymbol{Z};\widehat{\pi},\widehat{\boldsymbol{\mu}})]$, where $\mathbb{P}_n$ is the empirical measure over $\mathcal{D} = \{\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n\}$ and $(\widehat{\pi},\widehat{\boldsymbol{\mu}})$ is an estimate of $(\pi,\boldsymbol{\mu})$ from samples independent of $\mathcal{D}$.

Assume additionally the following holds for $j \in [p]$ with probability at least $1 - n^{-c}$ for some constant $c > 0$:

(1) Boundedness: $\widetilde{Y}_j$, $\mu_{aj}(\boldsymbol{W})$ and $\widehat{\mu}_{aj}(\boldsymbol{W})$ are bounded in $[-C, C]$ for some constant C.

(2) Bias of derived outcomes: $\Delta_{mj}(a) := \mathbb{E}[\widetilde{Y}_j(a) \mid \boldsymbol{W}, \boldsymbol{S}(a)] - Y_j(a)$ satisfies $\max_{j\in[p]} |\mathbb{E}[\Delta_{mj}(a)]| = \delta_m = o(n^{-1/2})$.

(3) Strict positivity: The true and estimated propensity score functions satisfy $\pi_a \geq \epsilon$ and $\widehat{\pi}_a \geq \epsilon$ for some constant $\epsilon \in (0, 1/2)$.

(4) Nuisance: The rates of nuisance estimates satisfy $\|\widehat{\pi}_a - \pi_a\|_{L_2} = \mathcal{O}(n^{-\alpha})$, $\max_{j\in[p]} \|\widehat{\mu}_{aj} - \mu_{aj}\|_{L_2} = \mathcal{O}(n^{-\beta})$ for some $\alpha, \beta \in (0, 1/2)$, and $\alpha + \beta > 1/2$.

Then as $m, n, p \to \infty$ such that $\log p = o(n^{\min(\frac{1}{2}, 2\alpha, 2\beta)})$, it holds that

(1) Asymptotic normality:

$$\sqrt{n}(\widehat{\tau}_{aj} - \tau_{aj}) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})\phi_{aj}(\boldsymbol{Z};\pi,\boldsymbol{\mu}) + \sqrt{n}\epsilon_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{aj}^2),$$

where $\sigma_{aj}^2 = \mathbb{V}[\phi_{aj}(\boldsymbol{Z};\pi,\boldsymbol{\mu})]$ and $\max_{j\in[p]} |\epsilon_j| = \mathcal{O}_{\mathbb{P}}((\log p)^{1/2} n^{-(1/2+\alpha\wedge\beta)} + \log p/n + n^{-(\alpha+\beta)} + \delta_m)$.

(2) Uniform control of variance estimation error: $\max_{j\in[p]} |\widehat{\sigma}_{aj}^2 - \sigma_{aj}^2| = \mathcal{O}_{\mathbb{P}}(n^{-\alpha\wedge\beta})$, where $\widehat{\sigma}_j^2 = \mathbb{V}_n[\phi_j(\boldsymbol{Z};\widehat{\mathbb{P}})]$.

Proof of Lemma B.4.1. We write $\widetilde{\tau}_{aj}(\mathbb{P}) = \mathbb{E}[\phi_{aj}(\boldsymbol{Z};\pi,\boldsymbol{\mu})]$. Note that $\tau_{aj} = \widetilde{\tau}_{aj} + \mathbb{E}[\Delta_{mj}(a)]$ from (B.3.1) in the proof of Proposition 10.

From Kennedy [84, Example 2], ϕ_{aj} is the efficient influence function of $\widetilde{\tau}_{aj}$. Then, from (3.2.2), we have a three-term decomposition

$$\begin{aligned}\widehat{\tau}_{aj}(\mathbb{P}) - \tau_{aj}(\mathbb{P}) &= \widehat{\tau}_{aj}(\mathbb{P}) - \widetilde{\tau}_{aj}(\mathbb{P}) + \widetilde{\tau}_{aj}(\mathbb{P}) - \tau_{aj}(\mathbb{P}) \\ &= T_{\mathrm{S},j} + T_{\mathrm{E},j} + T_{\mathrm{R},j} + \mathbb{E}[\Delta_{mj}], \end{aligned} \tag{B.4.7}$$

where

$$\begin{aligned} T_{\mathrm{S},j} &= (\mathbb{P}_n - \mathbb{P})\{\phi_{aj}(\boldsymbol{Z};\mathbb{P})\} \\ T_{\mathrm{E},j} &= (\mathbb{P}_n - \mathbb{P})\{\phi_{aj}(\boldsymbol{Z};\widehat{\mathbb{P}}) - \phi_{aj}(\boldsymbol{Z};\mathbb{P})\} \\ T_{\mathrm{R},j} &= \mathbb{P}\{\phi_{aj}(\boldsymbol{Z};\widehat{\mathbb{P}}) - \phi_{aj}(\boldsymbol{Z};\mathbb{P})\}. \end{aligned}$$

From assumption 2, we know that the last term can be uniformly controlled as $\delta_m = \max_{j\in[p]} |\mathbb{E}[\Delta_{mj}(a)]| = o_{\mathbb{P}}(n^{-1/2})$ as m, n, p tend to infinity.

Part (1) Uniform control of empirical process terms. We next verify that the conditions in Lemma 8 hold under the assumptions. Note that

$$\phi_{aj}(\boldsymbol{Z};\widehat{\mathbb{P}}) - \phi_{aj}(\boldsymbol{Z};\mathbb{P}) = \left(1 - \frac{\mathbb{1}\{A=a\}}{\pi_a(\boldsymbol{W})}\right)(\widehat{\mu}_{aj}(\boldsymbol{W}) - \mu_{aj}(\boldsymbol{W})) + \frac{\mathbb{1}\{A=a\}}{\widehat{\pi}_a(\boldsymbol{W})\pi_a(\boldsymbol{W})}(\widetilde{Y}_j - \widehat{\mu}_{aj}(\boldsymbol{W}))(\pi_a(\boldsymbol{W}) - \widehat{\pi}_a(\boldsymbol{W})).$$

Then, by the boundedness assumptions, we have that

$$\max_{j\in[p]} \|\phi_{aj}(\boldsymbol{Z};\widehat{\mathbb{P}}) - \phi_{aj}(\boldsymbol{Z};\mathbb{P})\|_{L_2} \leq \frac{1}{\epsilon}\max_{j\in[p]}\|\widehat{\mu}_{aj} - \mu_{aj}\|_{L_2} + \frac{1}{\epsilon^2}\|\widehat{\pi}_a - \pi_a\|_{L_2} = \mathcal{O}(n^{-\alpha\wedge\beta})$$
$$\left\|\max_{j\in[p]} |\phi_{aj}(\boldsymbol{Z};\widehat{\mathbb{P}}) - \phi_{aj}(\boldsymbol{Z};\mathbb{P})|\right\|_{L_\infty} \leq \frac{2C}{\epsilon^2},$$

From Lemma 8, when $\log p = o(n^{\min(\frac{1}{2},2\alpha,2\beta)})$, it follows that

$$\max_{j\in[p]} |T_{\mathrm{E},j}| = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}n^{-\alpha\wedge\beta} + \frac{\log p}{n}\right) = o\left(n^{-1/2}\right). \tag{B.4.8}$$

Part (2) Uniform control of remaining bias terms. From Kennedy [84, Example 54], it follows that

$$T_{\mathrm{R},j} = \mathbb{P}\left\{\left(\frac{1}{\pi_a} - \frac{1}{\widehat{\pi}_a}\right)(\mu_{aj} - \widehat{\mu}_{aj})\pi_a\right\}.$$

When $\widehat{\pi} \geq \epsilon$, conditioned on nuisance estimators, by the Cauchy-Schwarz inequality, we have

$$\max_{j\in[p]} |T_{\mathrm{R},j}| \leq \frac{1}{\epsilon}\|\widehat{\pi}_a - \pi_a\|_{L_2}\max_{j\in[p]}\|\widehat{\mu}_{aj} - \mu_{aj}\|_{L_2} = \mathcal{O}(n^{-(\alpha+\beta)}) = o(n^{-1/2}), \tag{B.4.9}$$

where the second inequality is from the Cauchy-Schwarz inequality, and the last equality is from the assumed rate for nuisance function estimation.

From (B.4.7)-(B.4.9), we have that

$$\widehat{\tau}_{aj}(\mathbb{P}) - \tau_{aj}(\mathbb{P}) = (\mathbb{P}_n - \mathbb{P})\{\phi_{aj}(\boldsymbol{Z};\mathbb{P})\} + \epsilon_j$$

such that $\max_{j\in[p]}|\epsilon_j| = \mathcal{O}((\log p)^{1/2}n^{-(1/2+\alpha\wedge\beta)} + \log p/n + n^{-(\alpha+\beta)} + \delta_m) = o(n^{-1/2})$.

Part (3) Sample average terms. By Lemma B.4.4, it follows that

$$\sqrt{n}T_{\mathrm{S},j} \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma_{aj}^2).$$

Part (4) Uniform control of variance estimates. We next verify that the conditions in Lemma 9 hold under the assumptions. Denote $\widehat{\phi}_j = \phi_j(\boldsymbol{Z};\widehat{\mathbb{P}})$ and $\phi_j = \phi_j(\boldsymbol{Z};\mathbb{P})$. Note that by the boundedness conditions, we have that

$$\max_{1\leq j\leq p}|\widehat{\phi}_j + \phi_j| \lesssim 1, \qquad \max_{1\leq j\leq p}|\widehat{\phi}_j - \phi_j| \lesssim 1, \tag{B.4.10}$$

which verifies the first condition of Lemma 9.

From Part (2), we also have

$$\max_{1\le j\le p}\|\widehat{\phi}_j-\phi_j\|_{L_1}\le\max_{1\le j\le p}\|\widehat{\phi}_j-\phi_j\|_{L_2}=\mathcal{O}\left(n^{-\alpha\wedge\beta}\right)$$

which verifies the last two conditions of Lemma 9.

Therefore, we are able to apply Lemma 9 to conclude that

$$\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|=\mathcal{O}\left(\frac{\log p}{n}+\sqrt{\frac{\log p}{n}}n^{-\alpha\wedge\beta}+n^{-\alpha\wedge\beta}\right)=\mathcal{O}(n^{-\alpha\wedge\beta}).$$

when $\log p=o(n^{\min(\frac{1}{2},2\alpha,2\beta)})$. □

Lemma B.4.2 (Error bounds of estimating equations)**.** Under the event that the conditions in Theorem 16 hold, it holds that

$$\max_{j\in[p]}|(\mathbb{P}_n-\mathbb{P})\{\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})\}|=\mathcal{O}\left(\sqrt{\frac{\log p}{n}}n^{-\alpha\wedge\beta}+\frac{\log p}{n}\right)$$
$$\max_{j\in[p]}|\mathbb{P}\{\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})\}|=\mathcal{O}\left(n^{-(\alpha+\beta)}\right).$$

Proof of Lemma B.4.2. We begin by decomposing the estimation error as

$$\begin{aligned}\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})&=\left(\frac{\mathbb{1}\{A=a\}}{\pi_a(\boldsymbol{W})}-1\right)(\widehat{\nu}_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}}))\\&\quad+\left(\frac{1}{\widehat{\pi}_a(\boldsymbol{W})}-\frac{1}{\pi_a(\boldsymbol{W})}\right)\mathbb{1}\{A=a\}(\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\psi(Y_j,\widehat{\theta}_{aj}^{\text{init}}))\\&\quad+\left(\frac{1}{\widehat{\pi}_a(\boldsymbol{W})}-\frac{1}{\pi_a(\boldsymbol{W})}\right)\mathbb{1}\{A=a\}(\widehat{\nu}_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}}))\\&=:D_{j1}(\boldsymbol{Z})+D_{j2}(\boldsymbol{Z})+D_{j3}(\boldsymbol{Z}).\end{aligned}$$

Next, we split the proof into two parts.

Part (1) From the boundedness assumption 1, we have that

$$\left\|\max_{j\in[p]}|\widehat{\omega}_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})|\right\|_{L_\infty}\le 2\left(\frac{1}{\epsilon}+1\right)+\frac{4}{\epsilon}+\frac{4}{\epsilon}\le C',$$

for some constant $C'>0$. This gives the L_∞-boundedness of the estimation error.

We next derive the upper bound of the L_2-norm by analyzing the three terms separately. Condition 2 implies that

$$\max_{j\in[p]}\|D_{j1}\|_{L_2}\le\frac{1}{\epsilon}\max_{j\in[p]}\sup_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)}\|\widehat{\nu}_{aj}(\boldsymbol{W},\theta)-\nu_{aj}(\boldsymbol{W},\theta)\|_{L_2}=\mathcal{O}(n^{-\alpha})$$
$$\max_{j\in[p]}\|D_{j2}\|_{L_2}\le\frac{1}{\epsilon^2}\|\widehat{\pi}_a-\pi_a\|_{L_2}=\mathcal{O}(n^{-\beta})$$
$$\max_{j\in[p]}\|D_{j3}\|_{L_2}\le\frac{1}{\epsilon^2}\|\widehat{\pi}_a-\pi_a\|_{L_2}=\mathcal{O}(n^{-\beta}).$$

Thus, we have

$$\max_{j\in[p]} \|\widehat{\omega}_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\|_{L_2} = \mathcal{O}(n^{-\alpha\wedge\beta}). \tag{B.4.11}$$

From Lemma 8, we have that

$$(\mathbb{P}_n - \mathbb{P})\{\widehat{\omega}_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\} = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} n^{-\alpha\wedge\beta} + \frac{\log p}{n}\right).$$

which finishes the proof of the first part of the lemma.

Part (2) On the other hand, because $\mathbb{E}[\mathbb{1}\{A = a\} \mid \boldsymbol{W}] = \pi_a(\boldsymbol{W})$, we have that

$$\mathbb{P}\{D_{j1}(\boldsymbol{Z})\} = 0.$$

Similarly, because $\mathbb{E}[\psi(Y_j, \widehat{\theta}_{aj}^{\text{init}}) \mid \boldsymbol{W}] = \nu_{aj}(\boldsymbol{W}, \widehat{\theta}_{aj}^{\text{init}})$, we also have that

$$\mathbb{P}\{D_{j2}(\boldsymbol{Z})\} = 0.$$

For the third term, we have that

$$\begin{aligned}\max_{j\in[p]} |\mathbb{P}\{D_{j3}(\boldsymbol{Z})\}| &\leq \frac{1}{\epsilon}\|\widehat{\pi}_a - \pi_a\|_{L_2} \max_{j\in[p]} \sup_{\theta\in\mathcal{B}(\widetilde{\theta}_{aj},\delta)} \|\widehat{\nu}_{aj}(\boldsymbol{W}, \theta) - \nu_{aj}(\boldsymbol{W}, \theta)\|_{L_2} \\ &= \mathcal{O}(n^{-(\alpha+\beta)}).\end{aligned}$$

Combining the above results, we have that

$$\max_{j\in[p]} |\mathbb{P}\{\widehat{\omega}_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\}| = \mathcal{O}(n^{-(\alpha+\beta)}).$$

□

Lemma B.4.3 (Error bounds of estimating equations evaluated with respect to the initial estimators)**.** Under the event that the conditions in Theorem 16 hold, it holds that

$$\begin{aligned}\max_{j\in[p]} |(\widehat{\theta}_{aj}^{\text{init}} - \widetilde{\theta}_{aj}) + f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})]| &= \mathcal{O}(n^{-2\gamma}) \\ \max_{j\in[p]} |(\mathbb{P}_n - \mathbb{P})[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})]| &= \mathcal{O}\left(\sqrt{\frac{\log p}{n}} n^{-\gamma/2} + \frac{\log p}{n}\right) \\ \max_{j\in[p]} |\mathbb{P}_n\{\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\}| &= \mathcal{O}\left(\frac{\log p}{n} + \sqrt{\frac{\log p}{n}} + n^{-\gamma}\right).\end{aligned}$$

Proof of Lemma B.4.3. We split the proof into different parts.

Part (1) By Taylor's expansion, we have that

$$\begin{aligned}\mathbb{P}\{\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})\} &= -\mathbb{P}\{\psi(\widetilde{Y}_j(a),\widehat{\theta}_{aj}^{\text{init}})\}\\ &= -f_{aj}(\widetilde{\theta}_{aj})(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj})+\mathcal{O}(|\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj}|^2)\\ &= -f_{aj}(\widetilde{\theta}_{aj})(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj})+\mathcal{O}(n^{-2\gamma})\\ &= \mathcal{O}(n^{-\gamma}),\end{aligned} \tag{B.4.12}$$

which is uniformly over $j\in[p]$ by Conditions 1 and 2. Note that

$$\begin{aligned}\mathbb{P}\{\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})\} &= -\mathbb{P}\psi\{(Y_j(a),\widehat{\theta}_{aj}^{\text{init}})\}+0\\ &= -\mathbb{P}\psi\{(Y_j(a),\widehat{\theta}_{aj}^{\text{init}})\}\\ &= -f_{aj}(\widetilde{\theta}_{aj})(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj})+\mathcal{O}(n^{-2\gamma}).\end{aligned}$$

Then, we have that

$$\begin{aligned}&\max_{j\in[p]}|(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj})+f_{aj}(\widetilde{\theta}_{aj})^{-1}\mathbb{P}[\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})]|\\ =&\max_{j\in[p]}|(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj})+f_{aj}(\widetilde{\theta}_{aj})^{-1}(-f(\widetilde{\theta}_{aj})(\widehat{\theta}_{aj}^{\text{init}}-\widetilde{\theta}_{aj}))|+\mathcal{O}(n^{-2\gamma})\\ =&\mathcal{O}(n^{-2\gamma}),\end{aligned}$$

which finishes the proof of the first part.

Part (2) Note that

$$\begin{aligned}&\max_{j\in[p]}\|\omega_{aj}(\boldsymbol{Z},\widehat{\theta}_{aj}^{\text{init}})-\omega_{aj}(\boldsymbol{Z},\widetilde{\theta}_{aj})\|_{L_2}\\ =&\max_{j\in[p]}\left[\left\|\left(\frac{\mathbb{1}\{A=a\}}{\pi_a(\boldsymbol{W})}-1\right)\left(\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\nu_{aj}(\boldsymbol{W},\widetilde{\theta}_{aj})\right)\right\|_{L_2}\right. \qquad\text{(B.4.13)}\\ &\left.+\left\|\frac{\mathbb{1}\{A=a\}}{\pi_a(\boldsymbol{W})}(\mathbb{1}\{\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\}-\mathbb{1}\{\widetilde{Y}_j\le\widetilde{\theta}_{aj}\})\right\|_{L_2}\right]\\ \lesssim&\max_{j\in[p]}\left[\left\|\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\nu_{aj}(\boldsymbol{W},\widetilde{\theta}_{aj})\right\|_{L_2}+\left\|\mathbb{1}\{\widetilde{Y}_j(a)\le\widehat{\theta}_{aj}^{\text{init}}\}-\mathbb{1}\{\widetilde{Y}_j(a)\le\widetilde{\theta}_{aj}\}\right\|_{L_2}\right]. \qquad\text{(B.4.14)}\end{aligned}$$

Since $\widehat{\theta}_{aj}^{\text{init}}$ is estimated from a separate independent sample, in the following analysis, we condition on $\widehat{\theta}_{aj}^{\text{init}}$. By Jenson's inequality, we have

$$\begin{aligned}&\mathbb{E}[(\nu_{aj}(\boldsymbol{W},\widehat{\theta}_{aj}^{\text{init}})-\nu_{aj}(\boldsymbol{W},\widetilde{\theta}_{aj}))^2]\\ =&\mathbb{E}[(\mathbb{P}(\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\mid\boldsymbol{W},A=a)-\mathbb{P}(\widetilde{Y}_j\le\widetilde{\theta}_{aj}\mid\boldsymbol{W},A=a))^2]\\ =&\mathbb{E}[(\mathbb{E}[\mathbb{1}\{\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\}-\mathbb{1}\{\widetilde{Y}_j\le\widetilde{\theta}_{aj}\}\mid\boldsymbol{W},A=a])^2]\\ \le&\mathbb{E}[\mathbb{E}[(\mathbb{1}\{\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\}-\mathbb{1}\{\widetilde{Y}_j\le\widetilde{\theta}_{aj}\})^2\mid\boldsymbol{W},A=a]]\\ =&\mathbb{E}[\mathbb{E}[\mathbb{1}\{\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\}+\mathbb{1}\{\widetilde{Y}_j\le\widetilde{\theta}_{aj}\}-2\,\mathbb{1}\{\widetilde{Y}_j\le\widehat{\theta}_{aj}^{\text{init}}\wedge\widetilde{\theta}_{aj}\}\mid\boldsymbol{W},A=a]]\\ =&\mathbb{P}(\widetilde{Y}_j(a)\le\widehat{\theta}_{aj}^{\text{init}})+\mathbb{P}(\widetilde{Y}_j(a)\le\widetilde{\theta}_{aj})-2\mathbb{P}(\widetilde{Y}_j(a)\le\widehat{\theta}_{aj}^{\text{init}}\wedge\widetilde{\theta}_{aj})\end{aligned}$$

where the last equation follows from the identification equation $\mathbb{E}[\mathbb{E}(\mathbb{1}\{\widetilde{Y}_j \le \widetilde{\theta}\} \mid \boldsymbol{W}, A=a)] = \mathbb{P}(\widetilde{Y}_j(a) \le \widetilde{\theta})$. Since f_{aj} is uniformly bounded in $\mathcal{B}(\widetilde{\theta}_{aj}, \delta)$ from Assumption 1, by Taylor's expansion we have

$$\begin{aligned}
&\mathbb{P}(\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}) + \mathbb{P}(\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}) - 2\mathbb{P}(\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}} \wedge \widetilde{\theta}_{aj}) \\
=& |\mathbb{P}(\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}) - \mathbb{P}(\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj})| \\
\lesssim& |\widehat{\theta}_{aj}^{\text{init}} - \widetilde{\theta}_{aj}| = \mathcal{O}(n^{-\gamma}).
\end{aligned}$$

The upper bound is uniform over $j \in [p]$ and we have

$$\max_{j\in[p]} \left\| \nu_{aj}(\boldsymbol{W}, \widehat{\theta}_{aj}^{\text{init}}) - \nu_{aj}(\boldsymbol{W}, \widetilde{\theta}_{aj}) \right\|_{L_2} = \mathcal{O}(n^{-\gamma/2}).$$

Similarly, we have

$$\begin{aligned}
&\left\| \mathbb{1}\{\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}\} - \mathbb{1}\{\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}\} \right\|_{L_2}^2 \\
=& \mathbb{E}[\mathbb{1}\{\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}\} + \mathbb{1}\{\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}\} - 2\,\mathbb{1}\{\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}\}\,\mathbb{1}\{\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}\}] \\
=& \mathbb{P}(\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}) + \mathbb{P}(\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}) - 2\mathbb{P}(\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}} \wedge \widetilde{\theta}_{aj}) \\
=& \mathcal{O}(n^{-\gamma})
\end{aligned}$$

$$\max_{1\le j\le p} \left\| \mathbb{1}\{\widetilde{Y}_j(a) \le \widehat{\theta}_{aj}^{\text{init}}\} - \mathbb{1}\{\widetilde{Y}_j(a) \le \widetilde{\theta}_{aj}\} \right\|_{L_2} = \mathcal{O}(n^{-\gamma/2}).$$

From Lemma 8, we have that

$$\max_{j\in[p]} |(\mathbb{P}_n - \mathbb{P})\{\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})\}| = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} n^{-\gamma/2} + \frac{\log p}{n} \right),$$

which finishes the proof of the second part.

Part (3) Note that

$$\mathbb{P}_n\{\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\} = (\mathbb{P}_n - \mathbb{P})[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}}) - \omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] + \mathbb{P}_n[\omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})] + \mathbb{P}[\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})].$$

Since ω is centered and bounded, by Lemma 8 we have

$$\max_{1\le j\le p} |(\mathbb{P}_n - \mathbb{P})[\omega_{aj}(\boldsymbol{Z}, \widetilde{\theta}_{aj})]| = \mathcal{O}\left(\frac{\log p}{n} + \sqrt{\frac{\log p}{n}} \right) = \mathcal{O}\left(\sqrt{\frac{\log p}{n}} \right).$$

Combining it with (B.4.12) and Part (2), we further have

$$\begin{aligned}
\max_{j\in[p]} |\mathbb{P}_n\{\omega_{aj}(\boldsymbol{Z}, \widehat{\theta}_{aj}^{\text{init}})\}| &= \mathcal{O}\left(\sqrt{\frac{\log p}{n}} n^{-\gamma/2} + \frac{\log p}{n} + \sqrt{\frac{\log p}{n}} + n^{-\gamma} \right) \\
&= \mathcal{O}\left(\frac{\log p}{n} + \sqrt{\frac{\log p}{n}} + n^{-\gamma} \right)
\end{aligned}$$

which completes the proof. □

Lemma B.4.4 (Lindeberg CLT for triangular array)**.** Let $m = m_n$ and $p = p_n$ be two sequences indexed by n. Consider the influence-function-based linear expansion for estimator $\widehat{\tau}_j$ of τ_j:

$$\sqrt{n}(\widehat{\tau}_j - \tau_j) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\varphi_{m_n j}\} + \epsilon_{m_n j}, \qquad j = 1, \ldots, p$$

where $\varphi_{m_n j}$ is the influence function that depends on m and the residual ϵ_j's satisfy that $\max_{j \in p_n} |\epsilon_{m_n j}| = o_{\mathbb{P}}(1)$ as $n \to \infty$. Let $B_n^2 = \sum_{i \in [n]} \mathbb{V}(\varphi_{m_n j}(\boldsymbol{Z}_i))$. Further assume that (i) there exists a constant $c > 0$, such that $\mathbb{V}(\varphi_{m_n j}(\boldsymbol{Z}_1)) \geq c$, and (ii) there exists a sequence $\{L_n\}_{n \in \mathbb{N}}$ such that $\max_{i \in [n]} |\varphi_{m_n j}(\boldsymbol{Z}_i)| \leq L_n$ and $L_n / B_n \to 0$, then

$$\sqrt{n} \frac{\widehat{\tau}_j - \tau_j}{\mathbb{V}\{\varphi_{m_n j}\}^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$$

Proof of Lemma B.4.4. Note that $\varphi_{m_n j}$ is the centered influence function such that $\mathbb{E}[\varphi_{m_n j}(\boldsymbol{Z})] = 0$. Let $X_{nk} = \varphi_{m_n j}(\boldsymbol{Z}_k)$. From assumption (ii) that $\max_{k \in [n]} |X_{nk}| \leq L_n$ and $L_n / B_n \to 0$, we have that, for any $\xi > 0$,

$$\lim_{n \to \infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \mathbb{E}\left[X_{nk}^2 \mathbb{1}\{|X_{nk}| \geq \xi B_n\}\right] = 0.$$

This verifies Lindeberg's condition for a triangular array of random variables. From Billingsley [18, Theorem 27.2], it follows that

$$\frac{n(\mathbb{P}_n - \mathbb{P})\{\varphi_{m_n j}\}}{B_n} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

as $n \to \infty$. Because $\boldsymbol{Z}_i$'s are identically distributed, we have $B_n^2 = n\mathbb{V}(\varphi_{m_n j})$. This implies that

$$\frac{\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\varphi_{m_n j}\}}{\mathbb{V}(\varphi_{m_n j})^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

as $n \to \infty$.

From assumption (i) that $\mathbb{V}(\varphi_{m_n j}) \geq c > 0$ and $\max_{j \in p_n} |\epsilon_{m_n j}| = o_{\mathbb{P}}(1)$, we further have

$$\max_{j \in p_n} \frac{|\epsilon_{m_n j}|}{\mathbb{V}(\varphi_{m_n j})^{1/2}} = o_{\mathbb{P}}(1)$$

as $n \to \infty$. Consequently, the conclusion follows. □

B.5 Multiple testing

B.5.1 Proof of Lemma 18

Proof of Lemma 18. We present the proof for $\tau_j = \tau_j^{\mathrm{STE}}$, and the proof for $\tau_j = \tau_j^{\mathrm{QTE}}$ follows similarly, which we omit for simplicity. From Theorem 12 and Proposition 14, we have the following expansion on the statistic:

$$\begin{aligned}\sqrt{n}\frac{\widehat{\tau}_j - \tau_j^*}{\widehat{\sigma}_j} &= \frac{\sqrt{n}}{\widehat{\sigma}_j}\frac{1}{n}\sum_{i=1}^n \varphi_{ij} + \frac{\sqrt{n}\epsilon_j}{\widehat{\sigma}_j}, \\ &= \frac{\sqrt{n}}{\sigma_j}\frac{1}{n}\sum_{i=1}^n \varphi_{ij} + \left(\frac{1}{\widehat{\sigma}_j} - \frac{1}{\sigma_j}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_{ij} + \frac{\sqrt{n}\epsilon_j}{\widehat{\sigma}_j}, \\ &= \frac{\sqrt{n}}{\sigma_j}\frac{1}{n}\sum_{i=1}^n \varphi_{ij} + \epsilon_j', \end{aligned} \tag{B.5.1}$$

where $\epsilon_j' = \left(\frac{1}{\widehat{\sigma}_j} - \frac{1}{\sigma_j}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_{ij} + \frac{\sqrt{n}\epsilon_j}{\widehat{\sigma}_j}$. Note that the covariance matrix of the true scaled influence function is given by

$$\mathrm{Cov}\left(\frac{1}{\sqrt{n}\sigma_j}\sum_{i=1}^n \boldsymbol{\varphi}_{i\mathcal{S}}\right) = \boldsymbol{D}_{\mathcal{S}}^{-1}\boldsymbol{E}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1},$$

where $\boldsymbol{E}_{\mathcal{S}} = \mathbb{E}[\boldsymbol{\varphi}_{i\mathcal{S}}\boldsymbol{\varphi}_{i\mathcal{S}}^\top]$ and $\boldsymbol{D}_{\mathcal{S}} = \mathrm{diag}((\widehat{\sigma}_j)_{j\in\mathcal{S}})$. The associated gaussian vector is defined as $\boldsymbol{g}_{0\mathcal{S}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_{\mathcal{S}}^{-1}\boldsymbol{E}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1})$. We first consider one-sided test problems with the pair of maximum statistics and the Gaussian vector based on linear expansion (B.5.1):

$$\overline{M}_{\mathcal{S}} = \max_{j\in\mathcal{S}} \sqrt{n}\frac{\widehat{\tau}_j - \tau_j^*}{\widehat{\sigma}_j}, \qquad W_{\mathcal{S}} = \max_{j\in\mathcal{S}}(g_{\mathcal{S}})_j,$$

and the pair based on the true influence functions and variances:

$$\overline{M}_{0\mathcal{S}} = \max_{j\in\mathcal{S}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varphi_{ij}}{\sigma_j}, \qquad W_{0\mathcal{S}} = \max_{j\in\mathcal{S}}(g_{0\mathcal{S}})_j.$$

We next show that the distribution function of $\overline{M}_{\mathcal{S}}$ can be approximated by $W_{\mathcal{S}}$ uniformly.

Step (1) Bounding difference between $W_{\mathcal{S}}$ and $W_{0\mathcal{S}}$. From Theorem J.1 of Chernozhukov et al. [30],

$$\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x |\mathbb{P}(W_{\mathcal{S}} > x) - \mathbb{P}(W_{0\mathcal{S}} > x \mid \{\boldsymbol{Z}_i\}_{i=1}^n)| \xrightarrow{\mathrm{p}} 0. \tag{B.5.2}$$

is implied if $\sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1} - \boldsymbol{D}_{\mathcal{S}}^{-1}\boldsymbol{E}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1}\|_{\max} = \mathcal{O}_{\mathbb{P}}(n^{-c})$ for some $c > 0$. Because $\max_{j\in\mathcal{S}}\sigma_j^2 \geq c > 0$, when $\log p = o(n^{2(\frac{1}{4}\wedge\alpha\wedge\beta)})$ and $\delta_m = o(n^{-1/2})$, we have that $\max_{j\in\mathcal{S}}|\widehat{\sigma}_j^{-1} - \sigma_j^{-1}| = \mathcal{O}_{\mathbb{P}}(r_\sigma)$ from the proof of Proposition 14, where $r_\sigma = n^{-\alpha\wedge\beta} + \sqrt{\log p/n}$. On the other

hand, (B.5.1) also implies that,

$$
\begin{aligned}
&\sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\widehat{\boldsymbol{E}}_{\mathcal{S}} - \boldsymbol{E}_{\mathcal{S}}\|_{\max} \\
&= \max_{k,\ell\in\mathcal{A}^*} |\widehat{E}_{k\ell} - E_{k\ell}| \\
&= \max_{k,\ell\in\mathcal{A}^*} \left| \frac{1}{n}\sum_{i=1}^{n} (\widehat{\varphi}_{ik}\widehat{\varphi}_{i\ell} - \mathbb{E}[\varphi_{ik}\varphi_{i\ell}]) \right| \\
&= \max_{k,\ell\in\mathcal{A}^*} \left| \frac{1}{n}\sum_{i=1}^{n} [\widehat{\varphi}_{ik}(\widehat{\varphi}_{i\ell} - \varphi_{i\ell}) + (\widehat{\varphi}_{ik} - \varphi_{ik})\varphi_{i\ell} + (\varphi_{ik}\varphi_{i\ell} - \mathbb{E}[\varphi_{ik}\varphi_{i\ell}])] \right| \\
&\leq \max_{k,\ell\in\mathcal{A}^*} \left| \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_{ik}(\widehat{\varphi}_{i\ell} - \varphi_{i\ell}) \right| + \max_{k,\ell\in\mathcal{A}^*} \left| \frac{1}{n}\sum_{i=1}^{n} (\widehat{\varphi}_{ik} - \varphi_{ik})\varphi_{i\ell} \right| + \max_{k,\ell\in\mathcal{A}^*} \left| \frac{1}{n}\sum_{i=1}^{n} (\varphi_{ik}\varphi_{i\ell} - \mathbb{E}[\varphi_{ik}\varphi_{i\ell}]) \right| \\
&\leq \max_{k,\ell\in\mathcal{A}^*} \max_{i} |\widehat{\varphi}_{ik}| \cdot \frac{1}{n}\sum_{i=1}^{n} |\widehat{\varphi}_{i\ell} - \varphi_{i\ell}| + \max_{k,\ell\in\mathcal{A}^*} \frac{1}{n}\sum_{i=1}^{n} |\widehat{\varphi}_{ik} - \varphi_{ik}| \cdot \max_{i} |\varphi_{i\ell}| + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right) \\
&= \mathcal{O}_{\mathbb{P}}(r_\varphi)
\end{aligned}
$$

where in the last inequality we use (B.4.4) with $r_\varphi = n^{-\alpha\wedge\beta} + \sqrt{\log p/n}$ and the sub-Gaussianity of $\varphi_{ik}\varphi_{i\ell}$'s. Then we have

$$
\begin{aligned}
\sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1} - \boldsymbol{D}_{\mathcal{S}}^{-1}\boldsymbol{E}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1}\|_{\max} &\leq \sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1} - \boldsymbol{D}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1}\|_{\max} \\
&\quad + \sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\boldsymbol{D}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1} - \boldsymbol{D}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1}\|_{\max} \\
&\quad + \sup_{\mathcal{S}\subseteq\mathcal{A}^*} \|\boldsymbol{D}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1} - \boldsymbol{D}_{\mathcal{S}}^{-1}\boldsymbol{E}_{\mathcal{S}}\boldsymbol{D}_{\mathcal{S}}^{-1}\|_{\max} \\
&= \mathcal{O}_{\mathbb{P}}(r_\varphi + r_\sigma).
\end{aligned}
$$

Under the condition that $\log^2(pn)\max\{\log^5(pn)/n, n^{-(\alpha\wedge\beta)}\} = o(1)$, we have that $(\log p)^2(r_\varphi + r_\sigma) = o(1)$. Then, from Theorem J.1 of Chernozhukov et al. [30], it follows that (B.5.2) holds.

Step (2) Bounding difference between $W_{0\mathcal{S}}$ and $\overline{M}_{0\mathcal{S}}$. Under Assumption 8 and the condition that $\log(pn)^7/n \leq C_2 n^{-c_2}$, because $\mathbb{E}[\varphi_{ij}^2]/\sigma_j^2 = 1$ and constant c, C is independent of $\mathcal{S}$, from Corollary 2.1 of Chernozhukov et al. [30], we have that

$$
\sup_{\mathcal{S}\subseteq\mathcal{A}^*} \sup_{x} |\mathbb{P}(W_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{0\mathcal{S}} > x)| \to 0. \tag{B.5.3}
$$

Step (3) Bounding difference between $\overline{M}_{0\mathcal{S}}$ and $\overline{M}_{\mathcal{S}}$. We begin by bounding $\max_{j\in\mathcal{S}} |\epsilon_j'|$.

Because σ_j is uniformly lower bounded away from zero, we have $|\widehat{\sigma}_j - \sigma_j| = |(\widehat{\sigma}_j^2 - \sigma_j^2)/(\widehat{\sigma}_j + \sigma_j)| \lesssim |\widehat{\sigma}_j^2 - \sigma_j^2|$, which implies that $\max_{j\in[p]} |\widehat{\sigma}_j - \sigma_j| \lesssim r_\sigma$ with probability tending to one from the proof of Proposition 14. From the proof of Theorem 12 and the boundedness assumptions,

we have

$$\begin{aligned}\max_{j\in\mathcal{S}}|\epsilon'_j| \le \max_{j\in\mathcal{A}^*}|\epsilon'_j| &\le \max_{j\in\mathcal{A}^*}\left|\frac{1}{\sqrt{n}}\sum_i \frac{\varphi_{ij}}{\sigma_j\widehat{\sigma}_j}(\sigma_j-\widehat{\sigma}_j)\right| + \max_{j\in\mathcal{A}^*}\left|\frac{\sqrt{n}\epsilon_j}{\widehat{\sigma}_j}\right| \\ &= \mathcal{O}_{\mathbb{P}}\left(\max_{j\in\mathcal{A}^*}|\sigma_j-\widehat{\sigma}_j|\max_{j\in\mathcal{A}^*}\left|\frac{1}{\sqrt{n}}\sum_i\varphi_{ij}\right| + \max_{j\in\mathcal{A}^*}|\sqrt{n}\epsilon_j|\right) \\ &= \mathcal{O}_{\mathbb{P}}\left(r_\sigma\sqrt{\log p} + n^{-(\alpha\wedge\beta)}\sqrt{\log p} + (\log p)/\sqrt{n} + n^{1/2-(\alpha+\beta)} + \sqrt{n}\delta_m\right) \\ &= \mathcal{O}_{\mathbb{P}}\left(n^{-(\alpha\wedge\beta)}\sqrt{\log p} + (\log p)/\sqrt{n} + n^{1/2-(\alpha+\beta)} + \sqrt{n}\delta_m\right) \\ &= o_{\mathbb{P}}(\xi_n),\end{aligned}$$

where $\xi_n = [n^{-(\alpha\wedge\beta)}\sqrt{\log p} + (\log p)/\sqrt{n} + n^{1/2-(\alpha+\beta)} + \sqrt{n}\delta_m]\log(n)$. Then,

$$\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\mathbb{P}(|\overline{M}_{0\mathcal{S}} - \overline{M}_{\mathcal{S}}| > \xi_n) \le \mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) \to 0. \tag{B.5.4}$$

Then we have

$$\begin{aligned}&|\mathbb{P}(\overline{M}_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{\mathcal{S}} > x)| \\ \le\ &\mathbb{P}(\overline{M}_{0\mathcal{S}} \le x, \overline{M}_{\mathcal{S}} > x) + \mathbb{P}(\overline{M}_{0\mathcal{S}} > x, \overline{M}_{\mathcal{S}} \le x) \\ \le\ &\mathbb{P}(\overline{M}_{0\mathcal{S}} > x-\xi_n, \overline{M}_{0\mathcal{S}} \le x) + \mathbb{P}(\overline{M}_{0\mathcal{S}} \le x+\xi_n, \overline{M}_{0\mathcal{S}} > x) + 2\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) \\ \le\ &\mathbb{P}(x-\xi_n < \overline{M}_{0\mathcal{S}} \le x+\xi_n) + 2\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) \\ \le\ &\mathbb{P}(x-\xi_n < W_{0\mathcal{S}} \le x+\xi_n) + 2\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) + 2\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(W_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{0\mathcal{S}} > x)|\end{aligned}$$

which implies that

$$\begin{aligned}&\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(\overline{M}_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{\mathcal{S}} > x)| \\ \le\ &\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x\mathbb{P}(x-\xi_n \le W_{0\mathcal{S}} \le x+\xi_n) + 2\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) \\ &\quad + 2\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(W_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{0\mathcal{S}} > x)| \\ \lesssim\ &\xi_n\sqrt{\log(pn)} + 2\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\epsilon'_j| > \xi_n\right) + 2\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(W_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{0\mathcal{S}} > x)|,\end{aligned}$$

where the second inequality is from the anti-concentration inequality [30, Lemma 2.1]. Under the condition that $\max\{\log(pn)^7/n, \log(pn)^2 n^{-(\alpha\wedge\beta)}, \sqrt{n\log(pn)}\delta_m\} \le C_2 n^{-c_2}$, we have $\xi_n\sqrt{\log(pn)} \lesssim n^{-c_2}\log n = o(1)$. This implies that

$$\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(\overline{M}_{0\mathcal{S}} > x) - \mathbb{P}(\overline{M}_{\mathcal{S}} > x)| \to 0. \tag{B.5.5}$$

Step (4) Combining results from previous steps. Finally, combining (B.5.2)-(B.5.5) yields that

$$\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(\overline{M}_{0\mathcal{S}}>x)-\mathbb{P}(W_{0\mathcal{S}}>x\mid\{\boldsymbol{Z}_i\}_{i=1}^n)|\xrightarrow{\mathrm{p}}0.$$

Analogously results also hold for $\overline{M}'_{\mathcal{S}}=\max_j -\sqrt{n}(\widehat{\tau}_j-\tau_j)/\widehat{\sigma}_j$, $W'_{\mathcal{S}}=\max_j -g_j$, $\overline{M}'_{0\mathcal{S}}=\max_j -n^{-1/2}\sum_{i=1}^n\varphi_{ij}/\sigma_j$ and $W'_{0\mathcal{S}}=\max_j -(g_0)_j$, by applying the same argument. Thus, we have

$$\begin{aligned}&\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(M_{\mathcal{S}}>x)-\mathbb{P}(\|\boldsymbol{g}_{\mathcal{S}}\|_\infty>x\mid\{\boldsymbol{Z}_i\}_{i=1}^n)|\\ \le&\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(\overline{M}_{\mathcal{S}}>x)-\mathbb{P}(W_{\mathcal{S}}>x\mid\{\boldsymbol{Z}_i\}_{i=1}^n)|\\ &+\sup_{\mathcal{S}\subseteq\mathcal{A}^*}\sup_x|\mathbb{P}(\overline{M}'_{\mathcal{S}}<x)-\mathbb{P}(W'_{\mathcal{S}}<x\mid\{\boldsymbol{Z}_i\}_{i=1}^n)|\xrightarrow{\mathrm{p}}0,\end{aligned}$$

which finishes the proof. □

B.5.2 Proof of Proposition 19

Proof of Proposition 19. We split the proof into different parts.

Part (1) Exact recovery of the active set. From the proof of Lemma 18, we have

$$\max_{1\le j\le p}|\widehat{\sigma}_j^2-\sigma_j^2|=\mathcal{O}_{\mathbb{P}}(r_\sigma). \tag{B.5.6}$$

Recall $\mathcal{A}^*=\{j\in[p]\mid\sigma_j^2\ge c_1\}$. From Assumption 8, $\min_{j\in\mathcal{A}^*}\sigma_j^2\ge c_1$ for some constant $c_1>0$.

To screen out noninformative coordinates, define c_n as in Proposition 19 for some constant $c>0$ and

$$\mathcal{A}_1=\{j\in[p]\mid\widehat{\sigma}_j^2\ge c_n\}$$

which is a random quantity because $\widehat{\sigma}_j^2$ is the empirical variance. Then we have

$$\mathbb{P}(\mathcal{A}^*\neq\mathcal{A}_1)=\mathbb{P}\left(\max_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\widehat{\sigma}_j^2\ge c_n\right)+\mathbb{P}\left(\min_{j\in\mathcal{A}^*\setminus\mathcal{A}_1}\widehat{\sigma}_j^2<c_n\right).$$

For the first term, we have

$$\begin{aligned}\mathbb{P}\left(\max_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\widehat{\sigma}_j^2\ge c_n\right)=&\mathbb{P}\left(\max_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\widehat{\sigma}_j^2-\min_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\sigma_j^2\ge c_n-\min_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\sigma_j^2\right)\\ \le&\mathbb{P}\left(\max_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}|\widehat{\sigma}_j^2-\sigma_j^2|\ge c_n-\min_{j\in\mathcal{A}_1\setminus\mathcal{A}^*}\sigma_j^2\right)\\ \le&\mathbb{P}\left(\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|\ge c_n-\max_{j\in\mathcal{A}^{*c}}\sigma_j^2\right)\to0,\end{aligned}$$

where we use the fact $c_n\to0, \min_{j\in\mathcal{A}^*}\sigma_j^2\ge c_1$ and $\max_{j\in[p]}|\sigma_j^2-\widehat{\sigma}_j^2|=\mathcal{O}_{\mathbb{P}}(r_\sigma)=o_{\mathbb{P}}(c_n)$ as $m,n,p\to\infty$, from Assumption 8 and (B.5.6).

For the second term, similarly we have

$$\begin{aligned}
&\mathbb{P}\left(\min_{j\in\mathcal{A}^*\setminus\mathcal{A}_1}\widehat{\sigma}_j^2 < c_n\right)\\
\leq&\mathbb{P}\left(\min_{j\in\mathcal{A}^*\setminus\mathcal{A}_1}\widehat{\sigma}_j^2 < c_n, \max_{j\in[p]}|\sigma_j^2-\widehat{\sigma}_j^2|\leq c_n\right)+\mathbb{P}\left(\max_{j\in[p]}|\sigma_j^2-\widehat{\sigma}_j^2|>c_n\right)\\
\leq&\mathbb{P}\left(\min_{j\in\mathcal{A}^*}\sigma_j^2 < 2c_n\right)+\mathbb{P}\left(\max_{j\in[p]}|\sigma_j^2-\widehat{\sigma}_j^2|>c_n\right)\to 0.
\end{aligned}$$

Thus, we have

$$\mathbb{P}(\mathcal{A}^*=\mathcal{A}_1)\to 1.$$

Part (2) FWER. From Part (1), the family-wise error rate satisfies that

$$\begin{aligned}
\text{FWER}&=\mathbb{P}(\widehat{\mathcal{A}}\cap\mathcal{V}^{*c}\neq\varnothing)\\
&=\mathbb{P}(\widehat{\mathcal{A}}\cap\mathcal{V}^{*c}\neq\varnothing,\mathcal{A}^*=\mathcal{A}_1)+o(1).
\end{aligned}$$

Recall the maximal statistic is defined as $M_1=\max_{j\in\widehat{\mathcal{A}}_1}|\sqrt{n}(\widehat{\tau}_j-\tau_j)/\widehat{\sigma}_j|$ and the null hypothesis is rejected only if $M_1>\widehat{q}_1(\alpha)$ where $\widehat{q}_1(\alpha)$ is the multiplier bootstrap quantile. From Lemma B.5.1, we have that

$$\begin{aligned}
\limsup\mathbb{P}(\widehat{\mathcal{A}}\cap\mathcal{V}^{*c}\neq\varnothing)&\leq\limsup\mathbb{P}\left(\max_{j\in\mathcal{A}^*}|\sqrt{n}(\widehat{\tau}_j-\tau_j^*)/\widehat{\sigma}_j|>\widehat{q}_1(\alpha),\mathcal{A}^*=\mathcal{A}_1\right)\\
&\leq\alpha.
\end{aligned}$$

Since when $\mathcal{A}^*=\mathcal{A}_1$, $\widehat{q}_1(\alpha)$ is also the bootstrap quantile of $\max_{j\in\mathcal{A}^*}|\sqrt{n}(\widehat{\tau}_j-\tau_j^*)/\widehat{\sigma}_j|$. Therefore, combining the above results yields that

$$\limsup\text{FWER}\leq\alpha,$$

which finishes the proof. □

B.5.3 Proof of Theorem 20

Proof of Theorem 20. We split the proof into different parts.

Part (1) FDPex. Recall $\mathcal{V}$ is the output of the step-down and augment processes, and M_j is the maximal statistic at step j. Let $\mathcal{V}_{\ell^*}$ be the set of discoveries returned by the step-down process. From Proposition 19 (1), the active set $\mathcal{A}_{\ell+1}\subseteq\mathcal{A}_\ell\subseteq\mathcal{A}^*$ for all $\ell=1,2,\ldots,\ell^*-1$ with probability tending to one. If $\mathcal{V}_{\ell^*}\cap\mathcal{V}^*\neq\varnothing$, then for the first $H_0^{(j)}$ generating false discoveries, from Lemma B.5.1 and Proposition 19, under the null hypothesis $H_0^{(j)}$, we always have

$$\limsup\text{FWER}=\limsup\mathbb{P}(M_j>\widehat{q}_j(\alpha))\leq\alpha$$

as $m,n,p\to\infty$. This shows that the step-down procedure controls the family-wise error rate. By Genovese and Wasserman [58, Theorem 1], it follows that the FDP after the augmentation step satisfies that

$$\limsup\mathbb{P}(\text{FDP}>c)\leq\alpha,$$

which finishes the proof of the first conclusion.

Part (2) Power. From the proof of Chang et al. [24, Corollary 1], the standard results on Gaussian maximum imply that $\max_{\ell=1,\ldots,\ell^*} \widehat{q}_\ell(\alpha) = C\sqrt{\log p} + o_{\mathbb{P}}(1)$ for some constant $C > 0$.

We next show that if there exists a $j_0 \in \mathcal{A}_{\ell^*} \cap \mathcal{V}^*$, then the proposed maximum test on $\mathcal{A}_{\ell^*}$ is able to reject the null hypothesis that $\tau_{j_0} = \tau^*_{j_0}$ for $\tau^*_{j_0} = 0$, which implies the power is converging to 1. Formally, Let $\widehat{q}_{\ell^*}(\alpha)$ be the corresponding estimated upper α quantile of the maximum statistic from the multiplier bootstrap procedure at ℓ^*-th step. Let E denote the event that stepdown process stops at ℓ^*-th step and by definition of ℓ^* we have $\mathbb{P}(E) = 1$. Notice that

$$
\begin{aligned}
&\mathbb{P}(\mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing) \\
=&\mathbb{P}(\mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing, E) \\
=&\mathbb{P}\left(\max_{j\in\mathcal{A}_{\ell^*}} \sqrt{n}\frac{|\widehat{\tau}_j|}{\widehat{\sigma}_j} > \widehat{q}_{\ell^*}(\alpha), \mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing, E\right) + \mathbb{P}\left(\max_{j\in\mathcal{A}_{\ell^*}} \sqrt{n}\frac{|\widehat{\tau}_j|}{\widehat{\sigma}_j} \leq \widehat{q}_{j_0}(\alpha), \mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing, E\right)
\end{aligned}
$$

By definition the stepdown process does not stop at ℓ^*-th step if $M_{\ell^*} = \max_{j\in\mathcal{A}_{\ell^*}} \sqrt{n}\frac{|\widehat{\tau}_j|}{\widehat{\sigma}_j} > \widehat{q}_{\ell^*}(\alpha)$. Hence, the first term is zero. For the second term, suppose $j_0 \in \mathcal{A}_{\ell^*} \cap \mathcal{V}^*$ and we have

$$
\begin{aligned}
&\mathbb{P}\left(\max_{j\in\mathcal{A}_{\ell^*}} \sqrt{n}\frac{|\widehat{\tau}_j|}{\widehat{\sigma}_j} \leq \widehat{q}_{\ell^*}(\alpha), \mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing\right) \\
\leq&\mathbb{P}\left(\sqrt{n}\frac{|\widehat{\tau}_{j_0}|}{\widehat{\sigma}_{j_0}} \leq \widehat{q}_{\ell^*}(\alpha), \mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing\right) \\
\leq&\mathbb{P}\left(\sqrt{n}\frac{\left|\widehat{\tau}_{j_0} - \tau^*_{j_0}\right|}{\widehat{\sigma}_{j_0}} \geq \sqrt{n}\frac{\left|\tau^*_{j_0}\right|}{\widehat{\sigma}_{j_0}} - \widehat{q}_{\ell^*}(\alpha)\right),
\end{aligned}
$$

Because $\left|\tau^*_{j_0}\right|/\widehat{\sigma}_{j_0} = \left|\tau^*_{j_0}\right|/\sigma_{j_0} \cdot (1 + o_{\mathbb{P}}(1)) \geq c(\log(p)/n)^{1/2}$ for large m and n, under the assumed minimal signal strength condition, by Lemma 18 we have

$$
\mathbb{P}\left(\sqrt{n}\frac{\left|\widehat{\tau}_{j_0} - \tau^*_{j_0}\right|}{\widehat{\sigma}_{j_0}} \geq \sqrt{n}\frac{\left|\tau^*_{j_0}\right|}{\widehat{\sigma}_{j_0}} - \widehat{q}_{\ell^*}(\alpha)\right) \to 0,
$$

as $m, n, p \to \infty$, when $c > C$. Therefore we have

$$
\mathbb{P}(\mathcal{A}_{\ell^*} \cap \mathcal{V}^* \neq \varnothing) \to 0,
$$

and

$$
\mathbb{P}(\mathcal{V}^* \subseteq \mathcal{V}_{\ell^*}) \to 1.
$$

Because the augmentation step only adds more discoveries, $\mathcal{V}_{\ell^*} \subset \mathcal{V}$, it does not decrease the power. Therefore, the final set of discoveries $\mathcal{V}$ has power 1 asymptotically as m, n, p tend to infinity. □

B.5.4 Helper lemmas

Lemma B.5.1 (Quantile estimation based on Gaussian approximation)**.** Suppose the conditions of Lemma 18 hold. For $\boldsymbol{g}_{\mathcal{S}} \sim \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1}\widehat{\boldsymbol{E}}_{\mathcal{S}}\widehat{\boldsymbol{D}}_{\mathcal{S}}^{-1})$, define the conditional α-quantile $q_{\mathcal{S}}(\alpha) := \inf\{t \in \mathbb{R} \mid \mathbb{P}(\|\boldsymbol{g}_{\mathcal{S}}\|_\infty \leq t \mid \{\boldsymbol{Z}_i\}_{i=1}^n) \geq \alpha\}$. As $m, n, p \to \infty$, it holds that

$$
\sup_{H_0^{\mathcal{S}}:\mathcal{S}\subseteq\mathcal{A}^*} \sup_{\alpha\in(0,1)} |\mathbb{P}(\overline{M}_{\mathcal{S}} \leq q_{\mathcal{S}}(\alpha) \mid \{\boldsymbol{Z}_i\}_{i=1}^n) - \alpha| \xrightarrow{\mathrm{p}} 0.
$$

Proof of Lemma B.5.1. From (B.5.4) in the proof of Lemma 18, we know that condition (14) in Chernozhukov et al. [30] holds for $\Delta_{1\mathcal{S}} = \max_{j\in[p]} |\overline{M}_{\mathcal{S}} - \overline{M}_{0\mathcal{S}}|$ uniformly over $\mathcal{S} \subseteq \mathcal{A}^*$. On the other hand, with probability tending to one,

$$\begin{aligned}
\Delta_{2\mathcal{S}} &\leq \Delta_{2\mathcal{A}^*} \\
&= \max_{j\in\mathcal{A}^*} \mathbb{P}_n[(\varphi_{ij}/\sigma_j - \widehat{\varphi}_{ij}/\widehat{\sigma}_j)^2] \\
&\leq \max_{j\in\mathcal{A}^*} \mathbb{P}_n[(\varphi_{ij} - \widehat{\varphi}_{ij})^2]/\sigma_j^2 + \max_{j\in\mathcal{A}^*} \mathbb{P}_n[\widehat{\varphi}_{ij}^2](1/\sigma_j - 1/\widehat{\sigma}_j)^2 \\
&\lesssim \max_{j\in\mathcal{A}^*} \mathbb{P}_n[(\varphi_{ij} - \widehat{\varphi}_{ij})^2] + \max_{j\in\mathcal{A}^*} (\sigma_j - \widehat{\sigma}_j)^2 \\
&\lesssim \max_{j\in\mathcal{A}^*} \mathbb{P}_n[|\varphi_{ij} - \widehat{\varphi}_{ij}|] + \max_{j\in\mathcal{A}^*} |\sigma_j - \widehat{\sigma}_j| \\
&= \mathcal{O}_{\mathbb{P}}\left(r_\varphi + r_\sigma\right),
\end{aligned}$$

which follows similarly as in the proof of Step (1) for Lemma 18. Under the conditions that $\max\{\log(pn)^7/n, \log(pn)^2 n^{-(\alpha\wedge\beta)}\} \leq Cn^{-c}$, we have

$$\sup_{\mathcal{S}\subseteq\mathcal{A}^*} \mathbb{P}(\log(pn)^2\Delta_{2\mathcal{S}} > n^{-c}) \leq \sup_{\mathcal{S}\subseteq\mathcal{A}^*} \mathbb{P}(\log(pn)^2\Delta_{2\mathcal{S}} > n^{-c}/\log(n)) \to 0,$$

which verifies condition (15) in Chernozhukov et al. [30]. From the proof of Corollary 3.1 in Chernozhukov et al. [30], the conclusion follows. □

B.6 Experiment details

B.6.1 Estimation of QTE

Initial estimator

The initial IPW estimator $\widehat{\theta}_{aj}^{\text{init}}$ of the quantile $Y_j(a)$ can be obtained by solving the following estimating equation:

$$\mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\widehat{\pi}_a(\boldsymbol{W})} \psi(Y_j, \theta) \right\} = 0$$

where ψ is defined in (3.4.4).

Counterfactual density estimation

Consider an unconfounded observational study with $(W, A, Y) \sim \mathcal{P}$, where A is binary and Y is continuous. Assume consistency, positivity, and exchangeability as usual. Below, we derive an identifying expression and estimator for the density of $Y(a)$ at a point.

Define $V := \mathbb{1}\{Y \leq y\}$. Noting that the density of $Y(a)$ is simply the derivative of $\mathbb{P}(Y(a) \leq y)$. Letting $p_{Y(a)}(y)$ and $p_Y(y)$ denote the densities of $Y(a)$ and Y respectively, we have

$$\begin{aligned}
p_{Y(a)}(y) &= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}\left(Y(a) \leq y\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{E}\left(V^a\right) = \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{E}\left\{\mathbb{E}\left(V^a \mid W\right)\right\} \\
&= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{E}\left\{\mathbb{E}\left(V^a \mid W, A=a\right)\right\} \\
&= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{E}\{\mathbb{E}(V \mid W, A=a)\} \\
&= \mathbb{E}\left\{ \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}(Y \leq y \mid W, A=a) \right\} \\
&= \mathbb{E}\left\{p_Y(y \mid W, A=a)\right\},
\end{aligned}$$

where the exchanging of integrals and derivatives is permitted by Leibniz's integral rule combined with the fact that V is bounded. As for estimation, we can reduce counterfactual density estimation to statistical density estimation. A natural doubly robust analog for this problem [88] that takes inspiration from one-step correction and kernel density estimation is given by

$$\widehat{p}_{Y(a)}(y) := \frac{1}{h} \mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\widehat{\pi}\left(a, \boldsymbol{W}\right)} \left(K\left(\frac{y-Y}{h}\right) - K\left(\frac{y-\widehat{\mu}\left(a, \boldsymbol{W}\right)}{h}\right)\right) + K\left(\frac{y-\widehat{\mu}\left(a, \boldsymbol{W}\right)}{h}\right) \right\}$$

where K is a kernel and h is the kernel smoothing bandwidth. Similar pseudo observations could be used with IPW [81, Remark 4] or plug-in style techniques as well, but we omit them for brevity.

SQTE

When comparing the quantile effects among genes with different scales, one can also consider the standardized quantile treatment effects (SQTE):

$$\tau_j^{\text{SQTE}_\varrho} = \frac{Q_\varrho(Y_j(1)) - Q_\varrho(Y_j(0))}{\text{IQR}(Y_j(0))}, \tag{B.6.1}$$

where for a random variable U, $Q_\varrho[U]$ denote the ϱ-quantile of random variable U, and $\mathrm{IQR}(U) = Q(0.75) - Q(0.25)$ denote the median and interquartile range of U with quantile function Q. Typically, when $\varrho = 0.5$, the ϱ-quantile equals to the median $Q_\varrho(U) = \mathrm{Med}(U)$, and we reveal the standardized median treatment effects $\tau_j^{\mathrm{SQTE}} = (\mathrm{Med}[Y_j(1)] - \mathrm{Med}[Y_j(0)])/\mathrm{IQR}(Y_j(0))$.

B.6.2 Extra experimental results

Simulation

Recall that λ_j is the mean of the couterfactuals $X_j(0)$ for $j \in [p]$ and $X_j(1)$ for $j \notin \mathcal{V}^*$. To generate the set of active genes $\mathcal{V}^*$, we draw a sample from a Multinomial with 200 trials and $p = 8000$ categories with probability

$$\mathrm{softmax}(\{\log(\mathrm{sd}(\exp(\lambda_j)))\}_{j\in[p]}).$$

where the sample standard deviation across cells i is used to estimate the above probability. The setup suggests that genes with higher variations are more likely to be active. For a maximum signal strength $\theta_{\max}$, we first draw a relative signal strength $r_j \sim \mathrm{Beta}(1, \beta_r)$ and set the final signal strength to be $s_j := \theta_{\max} r_j$.

Then, we consider two simulation scenarios for $X_j(1)$ with $j \in \mathcal{V}^*$ in the treatment group.

(1) Mean shift with high SNR

In this case, we set $(\theta_{\max}, \beta_r) = (1, 0.5)$ so that the more signals have magnitudes close to $\theta_{\max}$. Then, we adjust the effect sizes (λ_j) by adding or subtracting a signal (θ) with equal probability. Specifically, this can be represented as:

$$X_j(1) \sim \mathrm{Poisson}(\lambda_j + s_j\delta_j), \qquad j \in \mathcal{V}^*,$$

where $\delta_j \sim \mathrm{Bernoulli}(0.5)$.

(2) Median shift with low SNR

In this case, we set $(\theta_{\max}, \beta_r) = (10, 2)$ so that the more signals have magnitudes close to 0. Then we draw

$$X_j(1) \sim [\mathrm{LogNormal}(\lambda_j - s_j^2/2, s_j)], \qquad j \in \mathcal{V}^*,$$

which ensures that $X_j(1)$ has the same mean as $X_j(0)$, while their medians are different. Above, $[x]$ indicates rounding to ensure the generated expression levels are integers.

These DGPs aim to simulate complex data structures that reflect real-world phenomena, such as varying levels of signal perturbation and the impact of such variations on statistical analyses, particularly in the context of mean and median shifts in treated distributions.

We also inspect the effect of cross-fitting, which helps fulfill the sample splitting requirement and improves the estimation and inference accuracy. More specifically, we randomly split n observations into K disjoint folds $\mathcal{N}_1, \ldots, \mathcal{N}_K$. Then, for the kth fold $\mathcal{N}_k$, we compute the influence function values $\widehat{\varphi}_{ij}$ for $i \in \mathcal{N}_k$ with nuisance functions estimated from observations in other folds $\mathcal{N}_1, \ldots, \mathcal{N}_{k-1}, \mathcal{N}_{k+1}, \ldots \mathcal{N}_K$. We set the number of folds $K = 5$.

As shown in Figure B.62, the BH procedure for the ATE test controls the FDR at the desired level. When a large sample size, the FDR for the STE test with the BH procedure also gets closer to the desired threshold. This may be because the test statistics get more positively dependent on cross-fitting. However, the BH procedure does not control the FDR for the QTE test. In terms of FDX, the proposed multiple testing procedure gets tighter control, while the BH procedure

still fails to control it. Finally, the power deteriorates with cross-fitting compared to results in Figure 3.2.

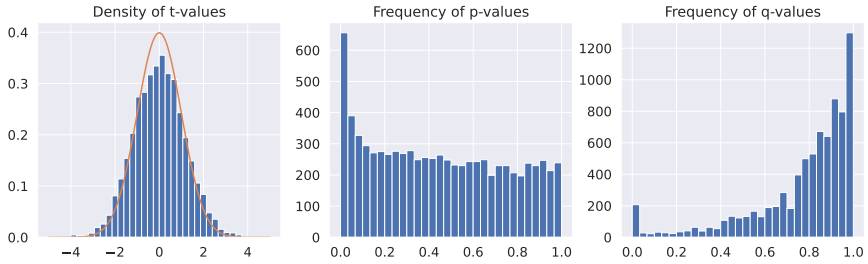

Figure B.61: The histogram of different statistics in one simulation of Figure 3.2 under mean shifts with $n = 100$. In this experiment, the number of true non-nulls is 200, while BH produces 258 discoveries with a q-value cutoff of 0.1, yielding 30% false discoveries.

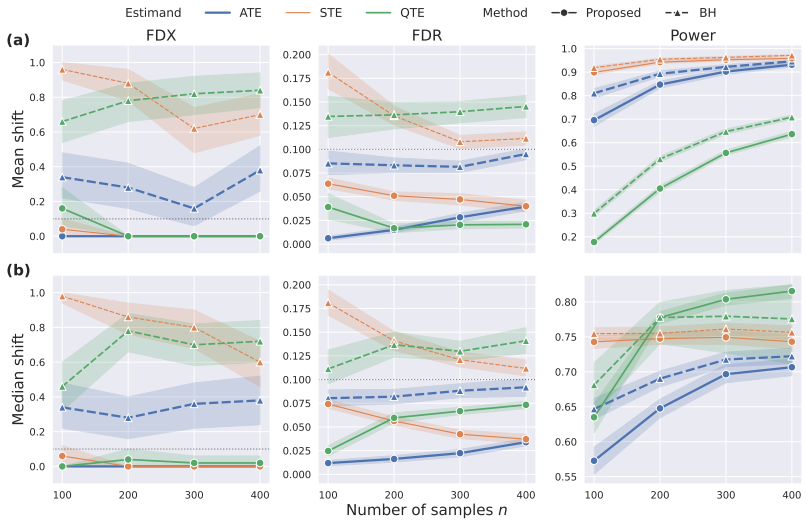

Figure B.62: Simulation results of the hypothesis testing of $p = 8000$ outcomes based on different causal estimands and FDP control methods for detecting differential signals under **(a)** mean shifts and **(b)** median shifts averaged over 50 randomly simulated datasets with 5-fold cross-fitting. The gray dotted lines denote the nominal level of 0.1.

LUHMES data

gRNA	Number of cells n	Number of genes p	Number of perturbed cells
PTEN	1014	1444	458
CHD2	756	1360	200
ASH1L	725	1360	169
ADNP	895	1416	339

Table B.61: The summary of sizes of data under different perturbations.

gRNA	Common	ATE only	STE only
PTEN	PTH2, PTGDS, NEFM, EEF1A1, C21orf59, MFAP4, ALCAM, NEFL, ITM2C, EIF3E, CRABP2, SLC25A6, EIF3L, WLS, PPP1R1C, GNB2L1, SVIP, RGS10, H3F3A, DRAXIN, GNG3, TCP10L, EIF3K	PCP4, MYL1, RAI14, DNER, MAP7, SNCA, TSC22D1, NRP2, SKIDA1	CCER2, PRDX1, TCF12
CHD2	EEF1A1, NEFL, GNG3, ID4, EEF2, STMN2	PRDX1, TUBB4B	PCP4
ASH1L	MT-CO1, MT-CYB, FXYD7		PKP4
ADNP	C21orf59, MAP1B	PTGDS, KLHL35, LHX2	

Table B.62: Sigfinicant genes for different guide RNA mutation on the late-stage cells. The last three columns show the discoveries that are significant in (1) both the ATE and the STE tests, (2) only the ATE tests, and (3) only the STE tests.

Lupus data

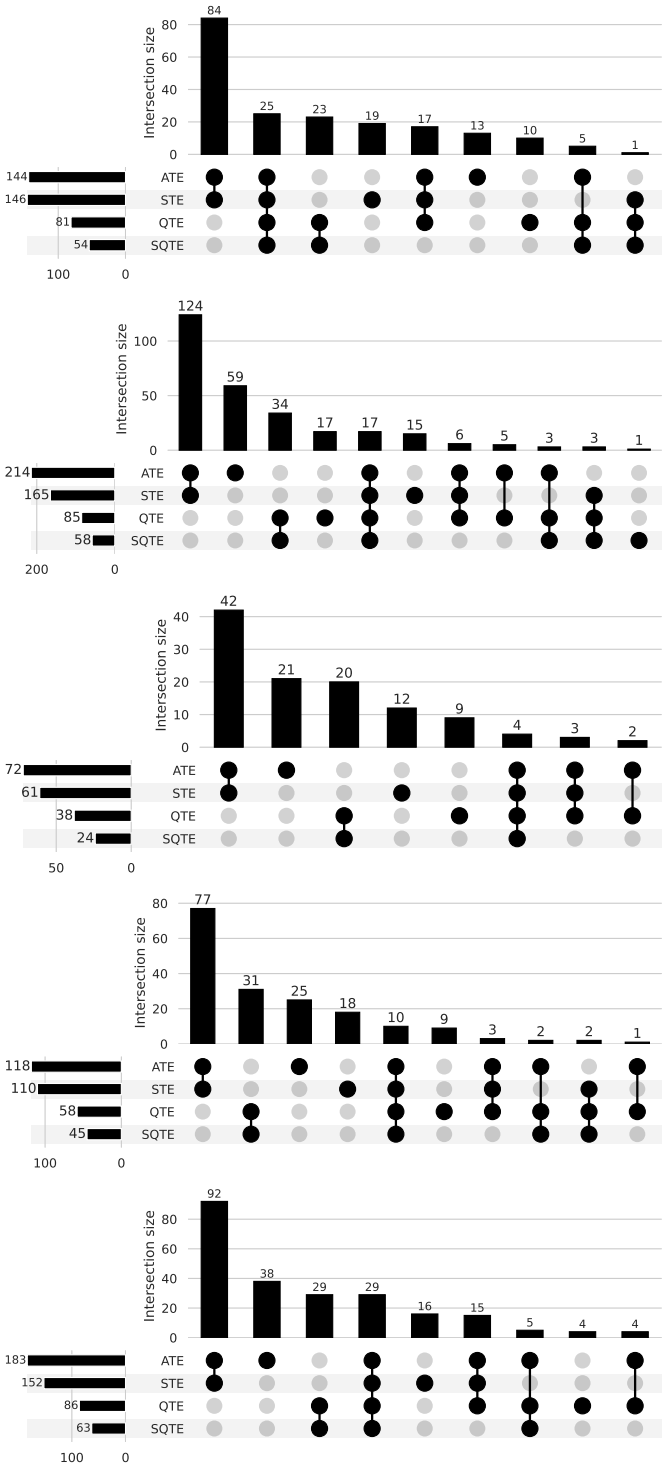

Figure B.63: Upset plot of discoveries by tests based on different causal estimands on the T4, T8, NK, B, and cM cell types of Lupus data set.

B.6.3 Perturb-seq data

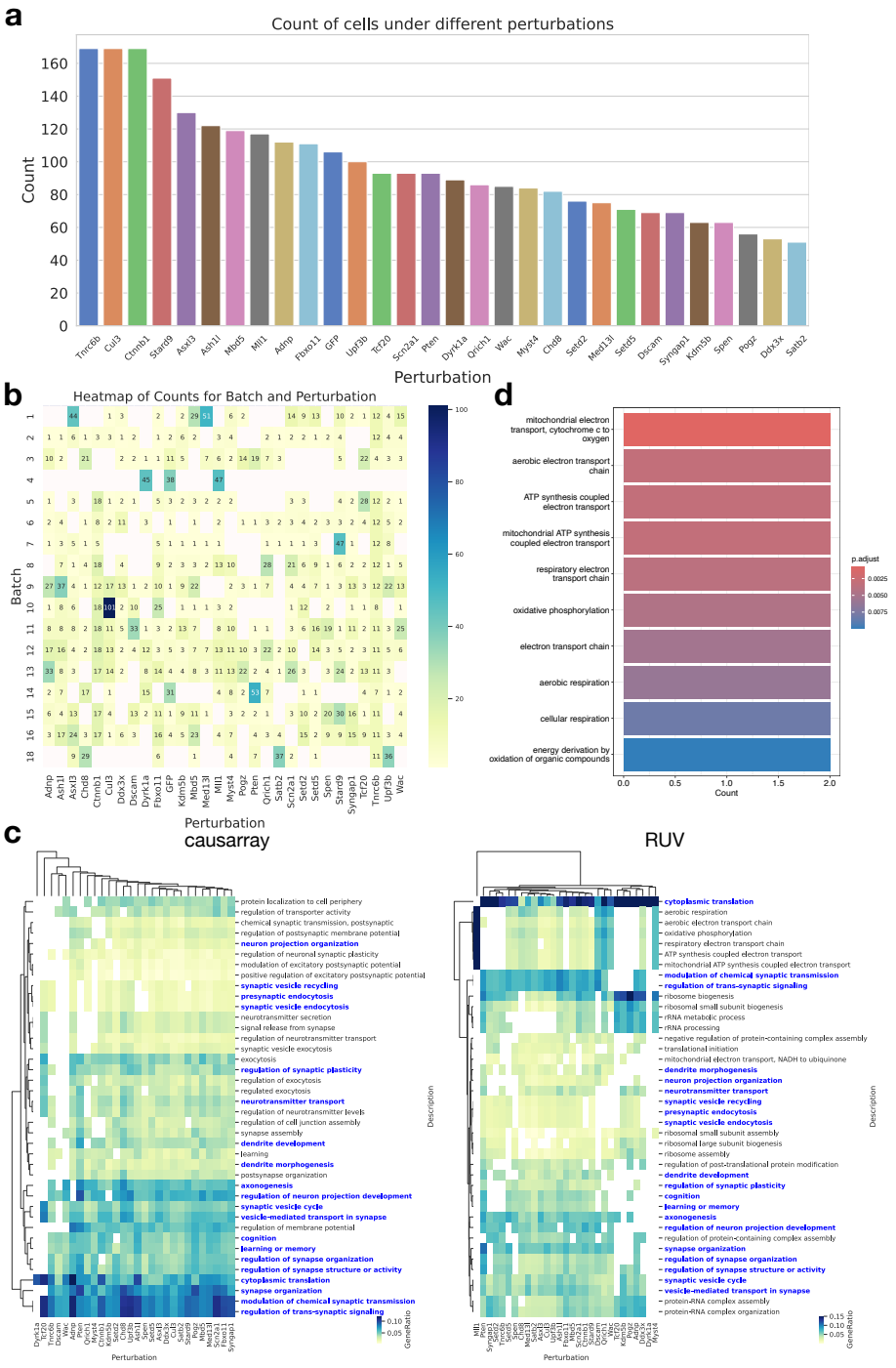

Figure B.64: Additional results on the Perturb-seq dataset. **a,** Barplot of the number of cells in each perturbation. **b,** Heatmap of the number of cells in each batch and perturbation. The batch design and the perturbation assignment of the Perturb-seq dataset are highly correlated. **c,** Clustermaps of GO terms enriched in discoveries (FDR<0.1) from causarray and RUV, respectively, where the common GO terms are highlighted in blue. Only the top 40 GO terms that have the most occurrences in all perturbations are displayed. **d,** Barplot of GO terms enriched in discoveries under *Mll1* perturbation from RUV.

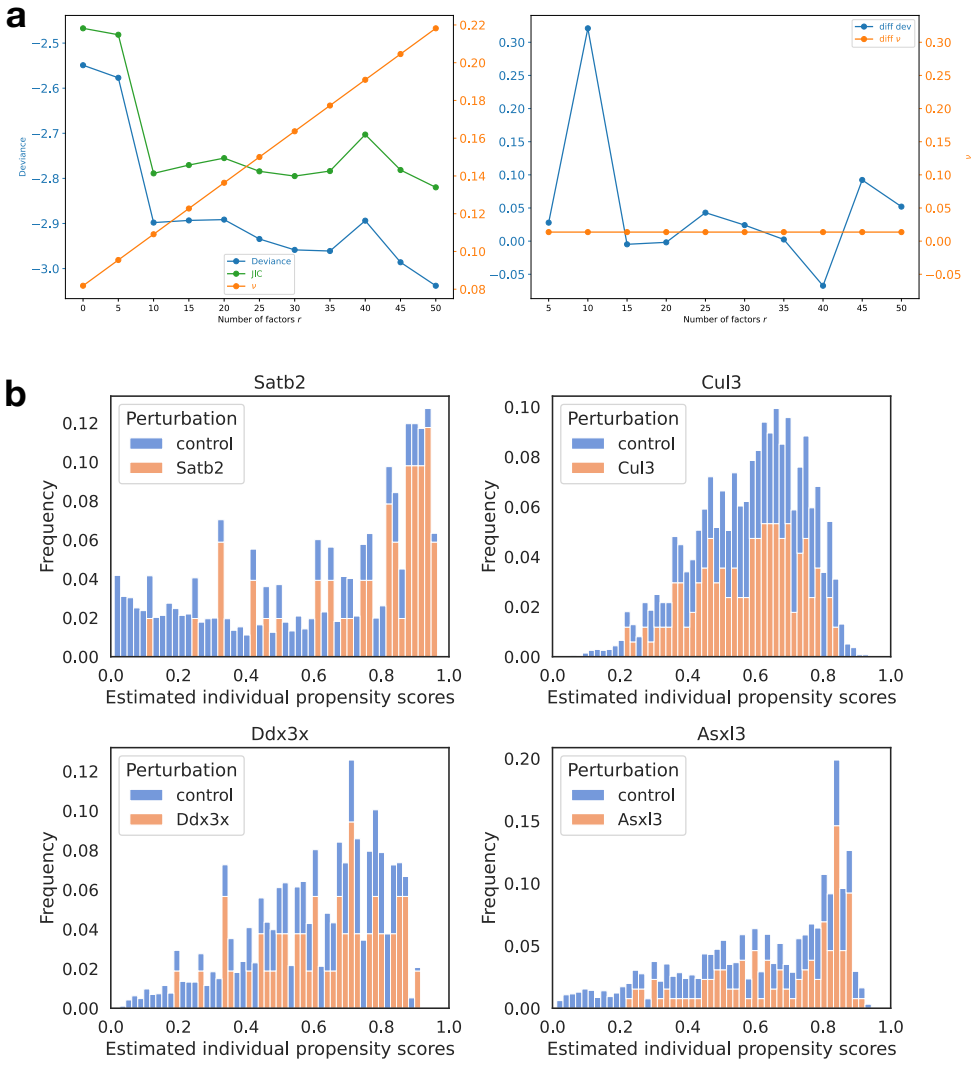

Figure B.65: Estimation results of causarray on the Perturb-seq dataset. **a,** The JIC criteria suggests a number of latent factor $r = 10$. **b,** Histograms of estimated propensity score for the top 4 perturbations (*Satb*2, *Cul*3, *Ddx*3*x*, and *Asxl*3) with most significant genes (adjusted P value < 0.1).

B.6.4 Alzheimer's data

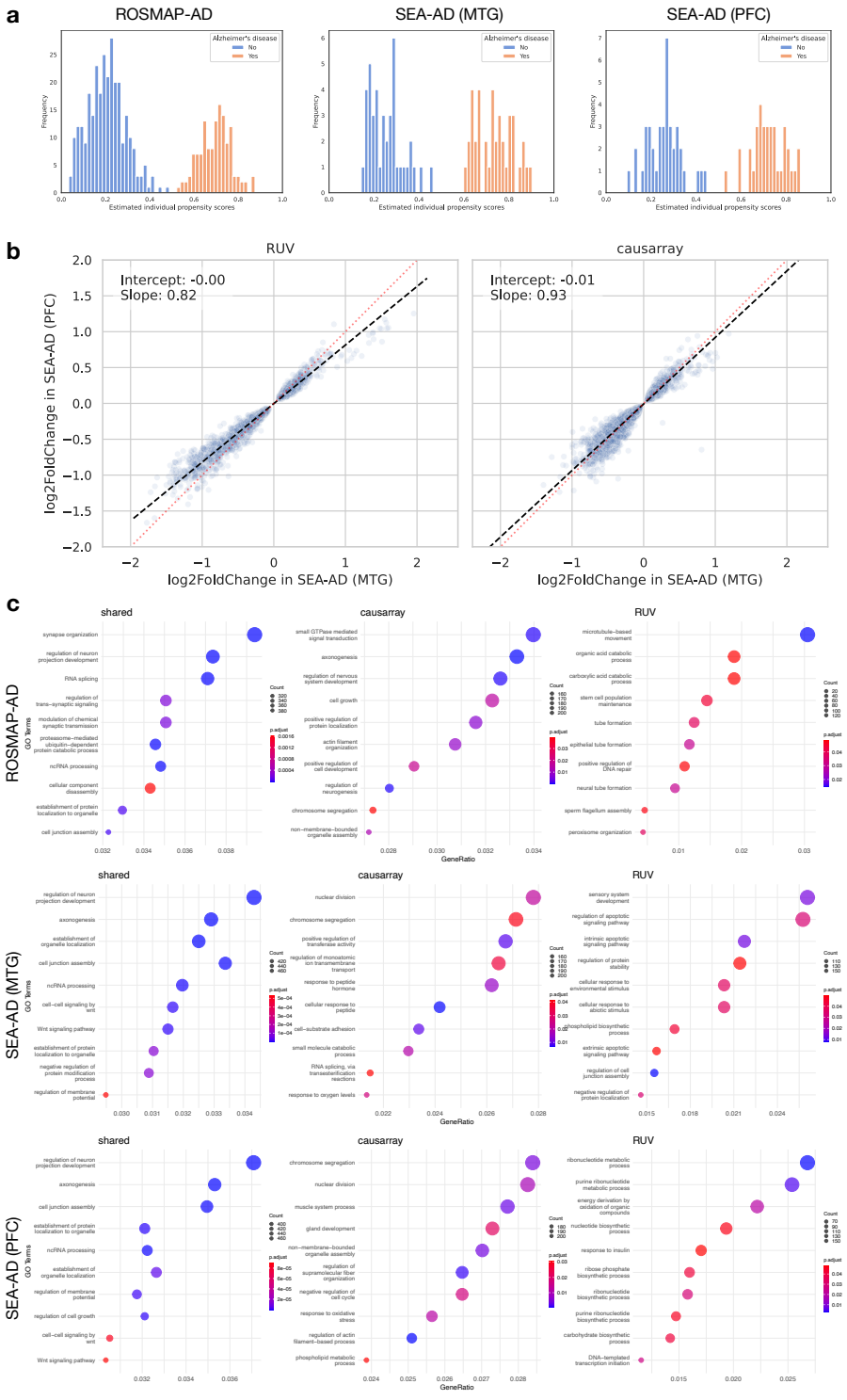

Figure B.66: Extra experimental results in AD datasets. a, Histogram of estimated propensity score in three AD datasets. **b,** Estimated effect sizes of DE genes (FDR < 0.001) in SEA-AD datasets. The black dashed line represents the fitted linear regression model, and the red dotted line represents the line $y = x$. **c,** Top gene ontology terms of the shared and distinct discoveries by causarray and RUV.

Appendix C

Assumption-Lean Post-Integrated Inference with Negative Control Outcomes

Notation. Throughout our exposition, we will use the following notational conventions. We use uppercase letters for random variables/vectors (e.g., Y, X, U) and lowercase for sample vectors, respectively (e.g., y, x, u). For a matrix $\beta \in \mathbb{R}^{d\times p}$, its jth column is denoted by $\beta_{\cdot j}$. Sets are denoted by calligraphic uppercase letters ($\mathcal{A}, \mathcal{C}$). Bold font is only used to denote design matrices and response matrices (e.g., $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{U}$) whose first dimension equals the sample size. For $p \in \mathbb{N}$, $[p] := \{1, \ldots, p\}$. For a set $\mathcal{A}$, let $|\mathcal{A}|$ be its cardinality.

For a random vector $X \in \mathbb{R}^p$, $\mathcal{P}_X$ denotes the projection in L_2. For any matrix $A \in \mathbb{R}^{n\times p}$ with full column rank, let $P_A = A(A^\top A)^{-1}A^\top$ and $P_A^\perp = I_p - P_A$ be the orthogonal projection matrices on the A's column space and its orthogonal space, respectively. For any square matrix $A \in \mathbb{R}^{n\times n}$, $\lambda_i(A)$ denotes its ith eigenvalue. The Gram matrix of $A^\top$ is denoted by $A^{\otimes 2} := AA^\top$. Matrix Hadamard product is denoted by $\odot$. For two symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, we write that $A \preceq B$ ($A \succeq B$) if $B - A$ ($A - B$) is positive semi-definite. For $a \in \mathbb{R}^m$, $\|a\|_q$ denotes the ℓ_q-norm for $q = 1, \ldots, \infty$. For $a \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, $\|a\|$ and $\|A\|$ denote the ℓ_2-norm and operator norm, respectively. The condition number of A is defined as $\kappa(A) = \|A\|\|A^{-1}\|$. For any random vector X, its L_q norm is defined as $\|X\|_{\mathrm{L_q}} = \mathbb{E}[\|X\|_q^q]^{1/q}$ for $q = 1, \ldots, \infty$.

For (potentially random) measurable functions f, we denote expectations with respect to Z alone by $\mathbb{P}f(Z) = \int f \,\mathrm{d}\mathbb{P}$, and with respect to both Z and the observations where f is fitted on by $\mathbb{E}[f(Z)]$. The empirical expectation is denoted by $\mathbb{P}_n f(Z) = \frac{1}{n}\sum_{i=1}^n f(Z_i)$. Similarly, the population and empirical variances (or covariance) are denoted by $\mathbb{V}$ and $\mathbb{V}_n$, respectively. The identity map is denoted by $\mathbb{I}$. We write the (conditional) L_p norm of f as $\|f\|_{\mathrm{L_p}} = \left[\int f(z)^p \,\mathrm{d}\mathbb{P}(z)\right]^{1/p}$ for $p \geq 1$.

We use "o" and "$\mathcal{O}$" to denote the little-o and big-O notations and let "$o_{\mathbb{P}}$" and "$\mathcal{O}_{\mathbb{P}}$" be their probabilistic counterparts. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ or $b_n \gg a_n$ if $a_n = o(b_n)$; $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if $a_n = \mathcal{O}(b_n)$; and $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Convergence in distribution and probability are denoted by "$\xrightarrow{\mathrm{d}}$" and "$\xrightarrow{\mathrm{p}}$". For $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

C.1 Related work

Batch correction and data integration. Large-scale single-cell transcriptomic datasets often include samples that span locations, laboratories, and conditions, leading to complex, nested batch effects in data [111, 165]. Batch correction specifically targets the removal of unwanted variation due to differences in batches within a single study, ensuring that the remaining data is comparable and reflects true biological differences. On the other hand, data integration focuses on combining and harmonizing multiple datasets to enhance statistical power and provide a more comprehensive analysis, dealing with both batch effects and between-dataset heterogeneity. Despite these differences, batch correction and data integration share the common goal of removing unwanted variation and preserving biological variation [186]. The integrated data is then used for downstream analysis, such as dimension reduction, clustering, and differential expressed gene testing. Integrated cellular profiles are typically represented as an integrated graph, a joint embedding, or a corrected response matrix. The main focus of the current paper is on the last category.

Despite the efforts from the computational biology and machine learning community to achieve better predictive power and data alignment, most existing batch correction methods are shown to be poorly calibrated [5, 114]. For statistical inference, many heuristic methods have been proposed to remove the batch effects and unwanted variations in the past decade. Leading examples include Remove Unwanted Variation (RUV) [56] and Surrogate Variable Analysis (SVA) [99]. RUV/SVA uses estimated factors of unwanted variation from unadjusted data, which works even if the batch design is unknown. When the batch design is known, two-step procedures for batch correction have also been proposed under parametric or mixture models [101, 112].

Unmeasured confounders adjustment and negative control outcomes. Over the past decades, researchers have been exploring methods to address the issue of unmeasured confounders in statistical analysis. In the presence of multiple outcomes, deconfounding techniques primarily employ two strategies: incorporating known negative control outcomes or leveraging sparsity assumptions [175]; while there is also another line of research on proximal causal inference, which uses both negative control outcomes and/or exposures for deconfounding [124]. For a comprehensive review of the literature on sparsity-based methods, readers are directed to Du et al. [48] and Zhou et al. [189]. This paper focuses on the negative control approach in the context of multiple outcomes.

Most existing works on confounder adjustment presume the knowledge of causal structure when the unobserved variable U is a mediator [175] and when U is a confounder Miao et al. [125], corresponding to Figure 4.3(a) and Figure 4.3(b), respectively. Recently developed sparsity-based methods by Bing et al. [20], Du et al. [48] have tried to relax this assumption to allow for a more flexible relationship between X and U. In particular, each entry of U can belong to different cases in Figure 4.3.

Negative control outcomes are used in observational studies under the key assumption that exposure has no causal effect on these outcomes. Rosenbaum [145] demonstrated that negative control outcomes can be employed to test for the presence of hidden confounding in observational studies. By introducing an additional variable known as a negative control exposure, Miao et al. [124] further showed that the average causal effect can be identified nonparametrically. Building upon this work, Shi et al. [156] developed a semiparametric inference procedure specifically for scenarios involving a categorical latent confounder and a binary exposure. Under linear latent models, Galbraith and Zinde-Walsh [57] use principal components of a set of potential controls

to adjust for unmeasured confounding effects. Under nonparametric models for a single outcome and multiple treatments Miao et al. [125] derive nonparametric identification conditions.

Assumption-lean semiparametric inference. There is increasing interest in deriving assumption-lean inference by using projection-based estimators [16] or semiparametric estimators [171]. The inferential problems we considered are also related to two-stage inference problems, such as post-sufficient dimension reduction inference [89], post-imputation inference [127], and inference with substituted covariate [2] or nonparametrically generated covariates [115]. While these related methods offer valuable insights into two-stage inference processes, they do not directly extend to address the challenges encountered in post-integrated inference problems.

C.2 Comparisons with related deconfounding approaches

C.2.1 Design-based approaches

As mentioned in the introduction, our paper mainly focuses on design-free data integration approaches. However, it is possible to relate the design-based approaches to design-free approaches so that the proposed method can be applied, as we discuss below. Design-based data integration approaches, such as Combat [80] and BUS [112], are usually based on a linear model:

$$Y_j = \alpha_j + X\beta_j + \gamma_{Bj} + \epsilon_{Bj}, \qquad j = 1, \ldots, p,$$

where $\alpha_j, \beta_j \in \mathbb{R}$ are coefficients for common variations while $\gamma_{Bj} \in \mathbb{R}$ is the location and $\epsilon_{Bj} \in \mathbb{R}$ is a mean-zero noise with scale differences across batches, respectively, for batch $B \in [n_B - 1]$ (with group $B = 0$ being the baseline and $\gamma_{0j} = 0$) and n_B is the total number of batches. This implies that

$$\mathbb{E}[Y_j \mid X, B] = \alpha_j + X\beta_j + \gamma_{Bj}, \qquad j = 1, \ldots, p.$$

Let $U_B \in \{0,1\}^{n_B}$ be the one-hot vector with only the B-th entry being one and zero elsewhere. Then, we can rewrite the above as

$$\mathbb{E}[Y \mid X, U_B] = \alpha + \beta^\top X + f(U_B)$$

where $f(U_B) = \gamma^\top U_B$ and $\gamma = [\gamma_{bj}]_{b\in[n_B], j\in[p]}$. In other words, the location-and-scale model considered by Johnson et al. [80] and Luo and Wei [112] is a special case of partial linear models with heterogenous noises, though they have utilized empirical Bayes shrinkage to improve the estimates. For this reason, a generalized least square approach could be used to improve Combat, as suggested by Li et al. [101].

In fact, when the additive noises are normal, we can decompose the noise as $\epsilon_{Bj} = U_\epsilon + Z_j$ for $B > 0$ such that $U_\epsilon \perp\!\!\!\perp \epsilon'_j$ and $\epsilon'_j \overset{\mathrm{d}}{=} \epsilon_{0j}$. To see this, define $\tau^2 = \mathbb{V}(\epsilon_{Bj})$ and $\sigma^2 = \mathbb{V}(\epsilon_{0j})$. Without loss of generality, we assume $\epsilon_{0j} \leq \min_{b\in[n_B]} \epsilon_{bj}$ so that $\tau^2 \geq \sigma^2$. If we define $U_\epsilon := \frac{1}{\tau(\tau^2-\sigma^2)}\epsilon_{Bj} + Z_j$ and $\epsilon'_j := \epsilon_{Bj} - U_\epsilon$, where $Z_j \sim \mathcal{N}(0, \frac{\sigma^2}{(\tau^2-\sigma^2)^2})$ is independent of ϵ_{Bj}, then U_ϵ and $\epsilon_{Bj} - U_\epsilon$ are independent because $\mathrm{Cov}(U_\epsilon, \epsilon_{Bj} - U_\epsilon) = (\tau^2(\tau^2-\sigma^2)^2)^{-1}\mathbb{V}(\epsilon_{Bj}) - \mathbb{V}(Z_j) = 0$. Here, we use the fact that two jointly normal random variables are independent if they are uncorrelated. In other words, we can rewrite the above model as

$$\mathbb{E}[Y \mid X, U] = \alpha + \beta^\top X + f(U),$$

where $f(U) = [\gamma, \mathbb{1}\{B > 0\}]^\top U$ and $U = [U_B, U_\epsilon]$. By absorbing part of the randomness of the additive noises into U, we convert the problem with heterogeneous noises into one with homogeneous noises studied in the current paper.

C.2.2 Unknown negative control outcomes

In this paper, we have focused on negative control outcomes to remove unwanted variations. When the negative control outcomes are unknown in advance, there are still possibilities to estimate the latent embedding and provide valid inferences. However, this typically requires extra sparsity assumptions on the effects of the covariate on multiple outcomes rather than utilizing the negative control outcomes. To illustrate the idea, we consider the following partial linear model:

$$\mathbb{E}[Y \mid X, U] = \beta^\top X + h(U).$$

Many methods start from the projected model

$$\mathbb{E}[\mathcal{P}_X^\perp Y \mid X, U] = \mathcal{P}_X^\perp h(U). \tag{C.2.1}$$

If the function $\mathcal{P}_X^\perp h$ has a good structure, then one may be able to recover U from $\mathcal{P}_X^\perp Y$. Alternatively, we can linearize the problem and seek partial recovery of the effect, as demonstrated below.

Example 1 (Linear models)**.** If h is a linear function such that $h : U \mapsto \eta^\top U$ for $\eta \in \mathbb{R}^{r\times p}$, then

$$Y = [\beta\ \eta]^\top \begin{bmatrix} X \\ U \end{bmatrix} + E.$$

With n i.i.d. samples, we obtain the following equation in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{U}\eta + \boldsymbol{E}. \tag{C.2.2}$$

Early methods in the literature rely on the assumption of the functional relationship between X and U. For example, Wang et al. [175] assume U to be a linear function of X with an additive Gaussian noise:

$$U = X\alpha + W, \tag{C.2.3}$$

in which case the sample counterpart of (C.2.1) reduces to

$$P_{\boldsymbol{X}}^\perp \boldsymbol{Y} = P_{\boldsymbol{X}}^\perp (\boldsymbol{W}\eta + \boldsymbol{E}). \tag{C.2.4}$$

Because the orthogonal projection is rank-deficient, one can further eliminate d rows of the above system of equations by elementary matrix transformation. For this purpose, Wang et al. [175] use QR decomposition by Householder rotation to derive a linear system of $n - d$ equations; e.g., Equation (2.5) and Equation (4.5) in Wang et al. [175] for $d = 1$ and $d > 1$, respectively. From this, $\widehat{\eta}$ is recovered from quasi-log-likelihood estimation. In the second step, the unknown coefficient (α, β) is estimated from (C.2.4) by plugging in the estimate $\widehat{\eta}$.

Under more general confounding machinism when (C.2.3) does not necessarily hold, Bing et al. [19, 20] rotate the original system to consistently estimate the marginal effect, under sparsity assumption on β and proper moment assumptions. They then use the residual from the lava

fit to uncover the column space of η. Finally, the partial coefficient $\beta P_\eta^\perp$ is recovered from the rotated system:

$$\boldsymbol{Y} P_\eta^\perp = \boldsymbol{X} \beta P_\eta^\perp + \boldsymbol{E} P_\eta^\perp.$$

These results have been extended to generalized linear models by Du et al. [48] using joint maximum likelihood estimation. When β is sparse, then it can be recovered by some estimator of $\beta P_\eta^\perp$ asymptotically. Note that the above approaches do not have too many restrictions on the observed covariate X and the latent embedding U, except for certain bounded moment assumptions.

Inspired by the success of methodologies development under linear models Example 1, one strategy for an extension to a nonlinear model is by linearizing the estimation problem. Specifically, suppose $h(U) = U\eta + R(U)$ for some remainder term R that depends on U, similarly we have a projection-based decomposition:

$$\begin{aligned} P_{\boldsymbol{X}}^\perp Y &= P_{\boldsymbol{X}}^\perp \boldsymbol{U}\eta + P_{\boldsymbol{X}}^\perp (R(\boldsymbol{U}) + \boldsymbol{E}) \\ \boldsymbol{Y} P_\eta^\perp &= \boldsymbol{X}\beta P_\eta^\perp + (R(\boldsymbol{U}) + \boldsymbol{E}) P_\eta^\perp, \end{aligned}$$

from which one may seamlessly use the methods by Bing et al. [19, 20] and Du et al. [48] when the remainder term can be well controlled.

Another possible strategy aligned with the angle of the current paper is to detect "weak" negative control outcomes and perform post-integrated inference based on such pseudo-negative control outcomes, as in Section 4.5. This approach is very similar to weak instrument detection and invalid instrumental variables selection; see, for example, Andrews et al. [4] and Windmeijer et al. [177]. We expect the rich literature on these related problems could lead to new methodological advances in post-integrated inference problems.

C.3 Nonparametric identification

Proof of Theorem 21. Under the equivalence assumption (Assumption 10 2), for any admissible distribution $\widetilde{f}(y_{\mathcal{C}}, u)$ we must have some invertible function v such that $\widetilde{f}(y_{\mathcal{C}}, u) = f\{Y_{\mathcal{C}} = y_{\mathcal{C}}, v(U) = u\}$. Note that (4.2.2) has at least one solution $\widetilde{f}(x \mid u) = f(x \mid v^{-1}(u))$; when this is the solution, define $\widetilde{f}(y_{\mathcal{C}^c} \mid x, u) := f(y_{\mathcal{C}^c} \mid x, v^{-1}(u))$. Then, $\widetilde{f}(y_{\mathcal{C}^c} \mid x, u)$ is also one solution to (4.2.3).

Because $v(U)$ is invertible, the ignorability assumption (Assumption 9 3) $Y(x) \perp\!\!\!\perp X \mid U$ implies that $Y(x) \perp\!\!\!\perp X \mid v(U)$; the completeness assumption Assumption 10 3 implies that $\widetilde{f}(u) > 0$ on $u \in v(\mathcal{U})$ and $\widetilde{f}(u \mid y_{\mathcal{C}}, x; \alpha)$ is also complete in $y_{\mathcal{C}}$. Further, from Assumption 9 2, the positivity condition $f_{X|v(U)}(x \mid v(u)) \in (0, 1)$ also holds for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. Then, we have

$$
\begin{aligned}
f_{Y(x)}(y) &= \int f_{Y(x)|U}(y \mid u) f(u) \, \mathrm{d}u \\
&= \int f_{Y(x)|U,X}(y \mid u, x) f(u) \, \mathrm{d}u && \text{(Assumption 9 2-3)} \\
&= \int f(y \mid u, x) f(u) \, \mathrm{d}u && \text{(Assumption 9 1)} \\
&= \int f(y_{\mathcal{C}^c} \mid u, x) f(y_{\mathcal{C}} \mid u) f(u) \, \mathrm{d}u && \text{(Assumption 10 1)} \\
&= \int f(y_{\mathcal{C}^c} \mid u, x) f(y_{\mathcal{C}}, u) \, \mathrm{d}u \\
&= \int \widetilde{f}(y_{\mathcal{C}^c} \mid x, u) \widetilde{f}(y_{\mathcal{C}}, u) \, \mathrm{d}u,
\end{aligned}
$$

where the last equality follows from the same derivation of g-formula applied on random variables $(Y, X, v(U))$. This completes the proof for the second conclusion.

We next show the uniqueness of the solutions to (4.2.2) and (4.2.3). For any candidate solutions $\widetilde{f}_1(x \mid u)$ and $\widetilde{f}_2(x \mid u)$ to (4.2.2), we must have that

$$
\int (\widetilde{f}_1(x \mid u) - \widetilde{f}_2(x \mid u)) \widetilde{f}(u) \, \mathrm{d}u = 0,
$$

which implies that $\widetilde{f}_1(x \mid U) - \widetilde{f}_2(x \mid U) = 0$ almost surely because of the completeness of $\widetilde{f}(u)$. Note that for any candidate solutions $\widetilde{f}_1(y_{\mathcal{C}^c} \mid x, u)$ and $\widetilde{f}_2(y_{\mathcal{C}^c} \mid x, u)$ to (4.2.3), we must have that

$$
\int (\widetilde{f}_1(y_{\mathcal{C}^c} \mid x, u) - \widetilde{f}_2(y_{\mathcal{C}^c} \mid x, u)) \widetilde{f}(u \mid y_{\mathcal{C}}, x) \, \mathrm{d}u \cdot f(y_{\mathcal{C}}, x) = 0.
$$

By the completeness property, this implies that $\widetilde{f}_1(y_{\mathcal{C}^c} \mid x, U) - \widetilde{f}_2(y_{\mathcal{C}^c} \mid x, U) = 0$ almost surely. Therefore, $\widetilde{f}(y_{\mathcal{C}^c} \mid x, u)$ is uniquely determined from (4.2.3). This completes the proof. □

C.4 Nonlinear main effects with estimated embeddings

C.4.1 Proof of Theorem 22

Proof of Theorem 22. Denote $A = \mathbb{E}[\mathrm{Cov}(X \mid U)]$, $\widehat{A} = \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U})]$, $B = \mathbb{E}[\mathrm{Cov}(X, \mathbb{E}[Y \mid X, U] \mid U)]$, and $\widehat{B} = \mathbb{E}[\mathrm{Cov}(X, \mathbb{E}[Y \mid X, U] \mid \widehat{U})]$. From Lemma C.4.1, we know that the error of two linear regression coefficients $\|\widetilde{\beta}_{\cdot j} - \beta_{\cdot j}\|$ is governed by $\|A - \widehat{A}\|$ and $\|B_{\cdot j} - \widehat{B}_{\cdot j}\|$, where the subscript j indicates the jth column of the corresponding matrices.

Part (1) Covariance estimation errors. To apply Lemma C.4.1, we first derive the error bounds for the two quantities. Note that $\mathrm{Cov}(X \mid U) = \mathbb{E}[X^{\otimes 2} \mid U] - \mathbb{E}[X \mid U]^{\otimes 2}$. We have

$$
\begin{aligned}
&\|A - \widehat{A}\| \\
=&\|\mathbb{E}[\mathrm{Cov}(X \mid U) - \mathrm{Cov}(X \mid \widehat{U})]\| \\
=&\|\mathbb{E}[\mathbb{E}[X \mid U]^{\otimes 2} - \mathbb{E}[X \mid \widehat{U}]^{\otimes 2}]\| \\
\le&\mathbb{E}[\|\mathbb{E}[X \mid U]^{\otimes 2} - \mathbb{E}[X \mid \widehat{U}]^{\otimes 2}\|] && \text{(Jensen's inequality)} \\
\le&\mathbb{E}[\|\mathbb{E}[X \mid U](\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}])^\top\| + \|(\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}])\mathbb{E}[X \mid \widehat{U}]^\top\|] \\
& && \text{(triangle inequality)} \\
=&\mathbb{E}[\|\mathbb{E}[X \mid U]\|\|\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}]\| + \|\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}]\|\|\mathbb{E}[X \mid \widehat{U}]\|] \\
\le&(\|\mathbb{E}[X \mid U]\|_{\mathrm{L}_2} + \|\mathbb{E}[X \mid \widehat{U}]\|_{\mathrm{L}_2})\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2} && \text{(Cauchy–Schwarz inequality)} \\
\le&2\|X\|_{\mathrm{L}_2}\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2}. && \text{(Jensen's inequality)}
\end{aligned}
$$

Similarly, the second covariance estimation error can be upper bounded as

$$
\begin{aligned}
&\|B_{\cdot j} - \widehat{B}_{\cdot j}\| \\
=&\|\mathbb{E}[\mathrm{Cov}(X, \mathbb{E}[Y_j \mid X, U] \mid U) - \mathrm{Cov}(X, \mathbb{E}[Y_j \mid X, \widehat{U}] \mid \widehat{U})]\| \\
=&\|\mathbb{E}[\mathbb{E}[X \mid U]\mathbb{E}[Y_j \mid X, U] - \mathbb{E}[X \mid \widehat{U}]\mathbb{E}[Y_j \mid X, \widehat{U}]]\| \\
\le&\mathbb{E}[\|\mathbb{E}[X \mid U]\mathbb{E}[Y_j \mid X, U] - \mathbb{E}[X \mid \widehat{U}]\mathbb{E}[Y_j \mid X, \widehat{U}]\|] && \text{(Jensen's inequality)} \\
\le&\mathbb{E}[\|\mathbb{E}[X \mid U](\mathbb{E}[Y_j \mid X, U] - \mathbb{E}[Y_j \mid X, \widehat{U}])\| + \|(\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}])\mathbb{E}[Y_j \mid X, \widehat{U}]\|] \\
& && \text{(triangle inequality)} \\
=&\mathbb{E}[\|\mathbb{E}[X \mid U]\|\|\mathbb{E}[Y_j \mid X, U] - \mathbb{E}[Y_j \mid X, \widehat{U}]| + \|\mathbb{E}[X \mid U] - \mathbb{E}[X \mid \widehat{U}]\|\|\mathbb{E}[Y_j \mid X, \widehat{U}]\|] \\
\le&\|\mathbb{E}[X \mid U]\|_{\mathrm{L}_2}\|\mathbb{E}[Y_j \mid X, \widehat{U}] - \mathbb{E}[Y_j \mid X, U]\|_{\mathrm{L}_2} + \|\mathbb{E}[Y_j \mid X, \widehat{U}]\|_{\mathrm{L}_2}\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2} \\
& && \text{(Cauchy–Schwarz inequality)} \\
\le&\|X\|_{\mathrm{L}_2}\|\mathbb{E}[Y_j \mid X, \widehat{U}] - \mathbb{E}[Y_j \mid X, U]\|_{\mathrm{L}_2} + \|Y_j\|_{\mathrm{L}_2}\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2}. && \text{(Jensen's inequality)}
\end{aligned}
$$

Part (2) Coefficient estimation error in terms of covariance estimation errors. When $\|\mathbb{E}[X \mid \widehat{U}] - \mathbb{E}[X \mid U]\|_{\mathrm{L}_2} < \sigma/(2M)$, from Part (1) we have $\kappa(A)\|A - \widehat{A}\|/\|A\| = \|A^{-1}\|\|A - \widehat{A}\| <$

1. From Lemma C.4.1, we further have that

$$\begin{aligned}\max_{j\in\mathcal{C}^c}\|\widetilde{\beta}_{\cdot j}-\beta_{\cdot j}\| &\le \frac{\max_{j\in\mathcal{C}^c}\kappa(A)\left(\frac{\|\beta_{\cdot j}\|\|A-\widehat{A}\|}{\|A\|}+\frac{\|B_j-\widehat{B}_j\|}{\|A\|}\right)}{1-\kappa(A)\frac{\|A-\widehat{A}\|}{\|A\|}}\\ &\lesssim \|X\|_{\mathrm{L}_2}\max_{j\in\mathcal{C}^c}\|\beta_{\cdot j}\|\|\mathbb{E}[Y_j\mid X,\widehat{U}]-\mathbb{E}[Y_j\mid X,U]\|_{\mathrm{L}_2}\\ &\quad+(\|X\|_{\mathrm{L}_2}+\max_{j\in\mathcal{C}^c}\|Y_j\|_{\mathrm{L}_2})\|\mathbb{E}[X\mid\widehat{U}]-\mathbb{E}[X\mid U]\|_{\mathrm{L}_2},\end{aligned}$$

where in the last inequality, we use the boundedness of A's spectrum from Assumption 12.

Part (3) Coefficient estimation error in terms of covariate estimation errors. From Lemma C.4.3, it follows that

$$\max_{j\in\mathcal{C}^c}\|\widetilde{\beta}_{\cdot j}-\beta_{\cdot j}\|\lesssim\left(\|X\|_{\mathrm{L}_2}(L_X^{\frac{1}{2}}+L_Y^{\frac{1}{2}})+\max_{j\in\mathcal{C}^c}\|Y_j\|_{\mathrm{L}_2}\right)\|v^{-1}(\widehat{U})-U)\|_{\mathrm{L}_2},$$

with $L_Y=\max_{j\in[p]}L_{Y_j}$. □

C.4.2 Proof of Lemma 23 (linear models)

Proof of Lemma 23. For observations $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{U})\in\mathbb{R}^{n\times d}\times\mathbb{R}^{n\times p}\times\mathbb{R}^{n\times r}$ and an estimate $\widehat{\boldsymbol{U}}\in\mathbb{R}^{n\times\widehat{r}}$ of $\boldsymbol{U}$, we have $S=\boldsymbol{X}^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}$, $\widetilde{S}=\boldsymbol{X}^\top P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{X}$, $\bar{\boldsymbol{Y}}=P_{\boldsymbol{U}}^\perp\boldsymbol{Y}$, and $\widetilde{\boldsymbol{Y}}=P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{Y}$. Furthermore, the regression coefficient on $(P_{\boldsymbol{U}}^\perp\boldsymbol{X},P_{\boldsymbol{U}}^\perp\boldsymbol{Y})$ can be expressed as

$$b=(\boldsymbol{X}^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X})^{-1}\boldsymbol{X}^\top P_{\boldsymbol{U}}^\perp\boldsymbol{Y}=S^{-1}\boldsymbol{X}^\top\bar{\boldsymbol{Y}}/n,$$

and the regression coefficient on $(P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{X},P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{Y})$ can be expressed as

$$\widetilde{b}=(\boldsymbol{X}^\top P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{X})^{-1}\boldsymbol{X}^\top P_{\widehat{\boldsymbol{U}}}^\perp\boldsymbol{Y}=\widetilde{S}^{-1}\boldsymbol{X}^\top\widetilde{\boldsymbol{Y}}/n.$$

From Lemma C.4.1, we have

$$\max_{j\in\mathcal{C}^c}\|\widetilde{b}_{\cdot j}-b_{\cdot j}\|\le\frac{\kappa(S)}{\|S\|_{\mathrm{op}}}\frac{\max_{j\in[p]}\|b_{\cdot j}\|\|\widetilde{S}-S\|+\|\boldsymbol{X}^\top(\widetilde{\boldsymbol{Y}}_{\cdot j}-\bar{\boldsymbol{Y}}_{\cdot j})/n\|}{1-\kappa(S)\frac{\|\widetilde{S}-S\|}{\|S\|_{\mathrm{op}}}}.\tag{C.4.1}$$

This requires verifying the assumptions therein. Specifically, we verify (C.4.4) below. Because

$$\begin{aligned}\|\widetilde{S}-S\|&=\|\boldsymbol{X}^\top(P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp)\boldsymbol{X}/n\|\\&\le\|\boldsymbol{X}^\top\boldsymbol{X}/n\|\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|\\&=\|S\|_{\mathrm{op}}\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|,\end{aligned}$$

and $\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|<1$ as assumed, we have

$$\frac{\|\widetilde{S}-S\|}{\|S\|_{\mathrm{op}}}<\frac{1}{\kappa(S)},$$

which verifies (C.4.4) of Lemma C.4.1. On the other hand, we also have

$$\|\boldsymbol{X}^\top(\widetilde{\boldsymbol{Y}}-\bar{\boldsymbol{Y}})\|_{2,\infty}=\|\boldsymbol{X}^\top(P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp)\boldsymbol{Y}\|_{2,\infty}\le\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|\|\boldsymbol{X}\|_{\mathrm{op}}\|\boldsymbol{Y}\|_{2,\infty}.$$

Therefore, (C.4.1) implies that

$$\begin{aligned}\max_{j\in\mathcal{C}^c}\|\widetilde{b}_{\cdot j}-b_{\cdot j}\|&\le\frac{\|S\|_{\mathrm{op}}\|b\|_{2,\infty}+\|\boldsymbol{X}\|_{\mathrm{op}}\|\boldsymbol{Y}\|_{2,\infty}/n}{\|S\|_{\mathrm{op}}}\frac{\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|}{1-\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|}\\&\le(\|b\|_{2,\infty}+\|S\|_{\mathrm{op}}^{-\frac12}\|\boldsymbol{Y}\|_{2,\infty}n^{-\frac12})\frac{\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|}{1-\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|}\end{aligned}$$

whenever $\|S\|_{\mathrm{op}}\neq0$ and $\kappa(S)\|P_{\widehat{\boldsymbol{U}}}^\perp-P_{\boldsymbol{U}}^\perp\|<1$. □

C.4.3 Auxillary lemmas

Lemma C.4.1 (Backward error of perturbed linear systems)**.** Let $A\in\mathbb{R}^{n\times n}$ be nonsingular, $b\in\mathbb{R}^n$, and $x=A^{-1}b\in\mathbb{R}^n$. In the following, $\Delta A\in\mathbb{R}^{n\times n}$ and $\Delta b\in\mathbb{R}^n$ are some arbitrary matrix and vector. We assume that the norm on A satisfies $\|Ax\|\le\|A\|\|x\|$ for all $A\in\mathbb{R}^{n\times n}$ and all $x\in\mathbb{R}^n$. Suppose $(A+\Delta A)\widehat{x}=\widehat{b}$ such that

$$\widehat{b}:=b+\Delta b\neq\mathbf{0} \tag{C.4.2}$$

$$\widehat{x}:=x+\Delta x\neq\mathbf{0} \tag{C.4.3}$$

$$\frac{\|\Delta A\|}{\|A\|}<\frac{1}{\kappa(A)}, \tag{C.4.4}$$

where $\kappa(A)=\|A\|\|A^{-1}\|$ is the condition number of A. Then, it holds that

$$\|\Delta x\|\le\frac{\|x\|\kappa(A)\frac{\|\Delta A\|}{\|A\|}+\kappa(A)\frac{\|\Delta b\|}{\|A\|}}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}.$$

If further, $b\neq\mathbf{0}$ (or equiavlently $x\neq\mathbf{0}$, then

$$\frac{\|\Delta x\|}{\|x\|}\le\frac{\kappa(A)\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right)}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}.$$

Proof of Lemma C.4.1. We split the proof into two parts.

Part (1) We first show that when (C.4.4) is satisfied, $A+\Delta A$ must be nonsingular. If $A+\Delta A$ is singular, then exists nonzero v such that $(A+\Delta A)v=\mathbf{0}$. Since A is nonsingular, we have $A^{-1}\Delta Av=-v$. So

$$\|v\|=\|A^{-1}\Delta Av\|\le\|A^{-1}\|\|\Delta A\|\|v\|,$$

which implies that

$$\|\Delta A\|\ge\frac{1}{\|A^{-1}\|}.$$

On the other hand, since $\kappa(A) = \|A\|\|A^{-1}\|$, from (C.4.4) we have

$$\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\|A\|\|A^{-1}\|},$$

or equivalently,

$$\|\Delta A\| < \frac{1}{\|A^{-1}\|}.$$

This leads to contradictions. Therefore, $A + \Delta A$ must be nonsingular.

Part (2) Since $(A + \Delta A)\widehat{x} = b + \Delta b$ and $Ax = b$, we have $A\Delta x + \Delta A\widehat{x} = \Delta b$. So $\Delta x = A^{-1}(\Delta b - \Delta A\widehat{x})$. Then we have

$$\begin{aligned}
\frac{\|\Delta x\|}{\|\widehat{x}\|} &= \frac{\|A^{-1}(\Delta b - \Delta A\widehat{x})\|}{\|\widehat{x}\|} \\
&\leq \frac{\|A^{-1}\|(\|\Delta A\|\|\widehat{x}\| + \|\Delta b\|)}{\|\widehat{x}\|} \\
&= \|A^{-1}\|\|A\| \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\|\|\widehat{x}\|} \right) \\
&= \kappa(A) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\|\|\widehat{x}\|} \right),
\end{aligned}$$

and

$$\begin{aligned}
\|\Delta x\| &\leq \kappa(A) \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\|\|\widehat{x}\|} \right) \|\widehat{x}\| \\
&= \kappa(A) \left(\frac{\|\Delta A\|}{\|A\|}\|\widehat{x}\| + \frac{\|\Delta b\|}{\|A\|} \right) \\
&\leq \kappa(A)\frac{\|\Delta A\|}{\|A\|}(\|x\| + \|\Delta x\|) + \kappa(A)\frac{\|\Delta b\|}{\|A\|}.
\end{aligned}$$

Rearrange the above inequality, we have

$$\begin{aligned}
\left(1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}\right)\|\Delta x\| &\leq \kappa(A)\frac{\|\Delta A\|}{\|A\|}\|x\| + \kappa(A)\frac{\|\Delta b\|}{\|A\|} \\
\|\Delta x\| &\leq \frac{\|x\|\kappa(A)\frac{\|\Delta A\|}{\|A\|} + \kappa(A)\frac{\|\Delta b\|}{\|A\|}}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}}.
\end{aligned}$$

When $x \neq \mathbf{0}$, we further have

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)\frac{\|\Delta A\|}{\|A\|} + \kappa(A)\frac{\|\Delta b\|}{\|A\|\|x\|}}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}} \leq \frac{\kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right)}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}},$$

where the last inequality holds since $\|b\| = \|Ax\| \leq \|A\|\|x\|$. □

Lemma C.4.2. Suppose X, Y are two random vectors in $\mathbb{R}^d$ defined on probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and $f : \mathbb{R}^d \to \mathbb{R}$ is a $\mathscr{F}$-measurable and satisfies the L-Lipschitz condition (in ℓ_q-norm) almost surely. Then it holds that

$$\|f(X) - f(Y)\|_{L_q} \leq L^{1/q} \|X - Y\|_{L_q}.$$

Proof of Lemma C.4.2. Note that

$$\begin{aligned}
\|f(X) - f(Y)\|_{L_q}^q &= \int |f(X) - f(Y)|^q \, \mathrm{d}\mathbb{P} \\
&\leq \int L \|X - Y\|_q^q \, \mathrm{d}\mathbb{P} && \text{(Lipschiz condition)} \\
&= L \sum_{j=1}^d \int |X_j - Y_j|^q \, \mathrm{d}\mathbb{P} \\
&= L \|X - Y\|_{L_q}^q
\end{aligned}$$

Then the conclusion follows by taking the q^{-1}-power on both sizes. □

Lemma C.4.3 (Error bound of regression function with estimated covariates)**.** On a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$, consider a random vector W and a sequence of random vectors $\{V_m\}_{m \in \mathbb{N}}$ adapted to a filtration $\{\mathscr{F}_m\}_{m \in \mathbb{N}}$ such that $\mathscr{F}_m \subseteq \mathscr{F}_{m+1}$. Suppose that (i) $\|W\|_{\mathrm{L}_2} < \infty$, (ii) $V_m \xrightarrow{\text{a.s.}} V$, and (iii) the function $h(v) = \mathbb{E}[W \mid V = v]$ satisfies the L-Lipschitz condition in ℓ_2-norm almost surely. Then, under (i)-(ii), it holds

$$\mathbb{E}[W \mid V_m] \xrightarrow{\mathrm{L}_2} \mathbb{E}[W \mid V],$$

and under (i)-(iii), it holds that

$$\|\mathbb{E}[W \mid V_m] - \mathbb{E}[W \mid V]\|_{\mathrm{L}_2} \leq 2L^{\frac{1}{2}} \|V_m - V\|_{\mathrm{L}_2}.$$

Proof of Lemma C.4.3. Define $\mathscr{F}_\infty = \sigma(\cup_m \mathscr{F}_m)$. There exists some $\mathscr{F}_\infty$-measurable function h and $\mathscr{F}_m$-measurable function h_m such that $h(V) = \mathbb{E}[W \mid V]$ and $h_m(V_m) = \mathbb{E}[W \mid V_m]$ almost surely. Notice that $(\mathbb{E}[W \mid V_m])_{m \in \mathbb{N}}$ is a Doob martingale (because $\|W\|_{\mathrm{L}_1} < \infty$). From martingale convergence theorem, there exists $V_\infty = \mathbb{E}[W \mid \mathscr{F}_\infty]$ that is measurable with respect to $\mathscr{F}_\infty$ such that $\|V_\infty\|_{\mathrm{L}_1} < \infty$ and $\mathbb{E}[W \mid V_m] \xrightarrow{\text{a.s.}} V_\infty$. On the other hand, because $V_m \xrightarrow{\text{a.s.}} V$ from Assumption (ii), we know that $V \overset{\text{a.s.}}{=} V_\infty$ is $\mathscr{F}_\infty$-measurable. This implies that $\mathbb{E}[W \mid V_\infty] = h(V_\infty) = h(V) = \mathbb{E}[W \mid V]$ almost surely. Thus, we conclude that $\mathbb{E}[W \mid V_m] \xrightarrow{\text{a.s.}} \mathbb{E}[W \mid V]$. By Jensen's inequality and Assumption (ii), we have $\|\mathbb{E}[W \mid V_m]\|_{\mathrm{L}_2} \leq \|W\|_{\mathrm{L}_2} < \infty$, which implies that the set of functions $\{\mathbb{E}[W \mid V_m] : m \in \mathbb{N}\}$ is uniformly integrable. Thus, we further have $\mathbb{E}[W \mid V_m] \xrightarrow{\mathrm{L}_2} \mathbb{E}[W \mid V]$ from dominated convergence theorem.

Next, we need to derive the convergence rate. We have that

$$\|h_m(V_m) - h(V)\|_{\mathrm{L}_2} \leq \|h_m(V_m) - h(V_m)\|_{\mathrm{L}_2} + \|h(V_m) - h(V)\|_{\mathrm{L}_2}. \tag{C.4.5}$$

For the first term in (C.4.5), from the martingale property, the function representation $h_m(V_m) = \mathbb{E}[h(V) \mid \mathscr{F}_m]$ gives that

$$\|h_m(V_m) - h(V_m)\|_{\mathrm{L}_2} = \|\mathbb{E}[h(V) \mid \mathscr{F}_m] - h(V_m)\|_{\mathrm{L}_2} \leq \|h(V) - h(V_m)\|_{\mathrm{L}_2} \tag{C.4.6}$$

where the last inequality is from Jensen's inequality.

Combining (C.4.5) and (C.4.6) yields that

$$\|h_m(V_m) - h(V)\|_{\mathrm{L}_2} \leq 2\|h(V_m) - h(V)\|_{\mathrm{L}_2} \leq 2L^{\frac{1}{2}}\|V_m - V\|_{\mathrm{L}_2},$$

where the last inequality is from Lemma C.4.2 by noting that h satisfies the L-Lipschitz condition from Assumption (iii). □

C.5 Doubly robust semiparametric inference

C.5.1 Proof of Theorem 24 and Corollary 25

Proof of Theorem 24 and Corollary 25. Theorem 24 is a special case of Theorem C.5.1 with nonlinear link functions. The proof follows by applying Theorem C.5.1 with g being identity. Meanwhile, the assumption in Theorem C.5.1 can be relaxed under this special case by noting that $\mathbb{E}[Y \mid X, U]$ can be replaced by Y because the residual $Y - \mathbb{E}[Y \mid X, U]$ is orthogonal to mean-zero functions of (X, U) in the L_2 space so that $\eta(O) = Y - \mathbb{E}[Y \mid U]$ under identity link. □

C.5.2 Proof of Proposition 26

Proof of Proposition 26. From Theorem 24, we have

$$\sqrt{n}(\widetilde{b} - \widetilde{\beta}) = \sqrt{n}\widetilde{\Sigma}^{-1}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(O; \mathbb{P})\} + \xi,$$

where the remainder term ξ satisfies that $\|\xi\|_{2,\infty} = o_{\mathbb{P}}(1)$ under the rate conditions in Theorem 24. Recall that $t_j = \sqrt{n}\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}^{\frac{1}{2}}\widehat{\Sigma}^{-1}(\widetilde{b}_{\cdot j} - \widetilde{\beta}_{\cdot j})$. We have that

$$\begin{aligned} v^\top t_j &= \sqrt{n}v^\top \mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\} \\ &\quad + (\sqrt{n}v^\top(\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}^{\frac{1}{2}} - \mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{\frac{1}{2}})(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\} + v^\top \mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}^{\frac{1}{2}}\widehat{\Sigma}^{-1}\xi_{\cdot j}) \\ &=: \vartheta_j + \varsigma_j. \end{aligned}$$

For the first component, note that ϑ_j for $j = 1, \ldots, p$ are independent conditional on (X, U)'s. Furthermore, the self-normalizing term $\vartheta_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$ for $j \in \mathcal{N}_p$. By the strong law of large number, $\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\} \xrightarrow{\text{a.s.}} \mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}$, $\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\} \xrightarrow{\text{a.s.}} \mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}$ and $\widehat{\Sigma} \xrightarrow{\text{a.s.}} \widetilde{\Sigma}$ when both $m, n, p \to \infty$. This implies that $\max_j \|\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\} - \mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\|_{\text{op}} \xrightarrow{\text{a.s.}} 0$, $\max_j \|\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\} - \mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\|_{\text{op}} \xrightarrow{\text{a.s.}} 0$, $\|\widetilde{\Sigma} - \widehat{\Sigma}\|_{\text{op}} \xrightarrow{\text{a.s.}} 0$, which follows from Patil et al. [136, Lemma S.8.6 (1)] by noting that the variables O's are iid in the triangluar array. For the second component, we have that

$$\begin{aligned} \max_{1 \le j \le p} |\varsigma_j| &\le \max_{1 \le j \le p} \|\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}^{\frac{1}{2}} - \mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{\frac{1}{2}}\|_{\text{op}} \cdot \|\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\|_2 \\ &\quad + \max_{1 \le j \le p} \|\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}^{\frac{1}{2}}\widehat{\Sigma}^{-1}\|_{\text{op}}\|\xi_{\cdot j}\|_2 \\ &= o_{\mathbb{P}}(1)\mathcal{O}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1)o_{\mathbb{P}}(1) \\ &= o_{\mathbb{P}}(1). \end{aligned}$$

Let $\varrho = |\mathcal{N}_p|^{-1}\sum_{j \in \mathcal{N}_p} \mathbb{1}\{|v^\top t_j| > z_{\frac{\alpha}{2}}\}$. To prove the overall Type-I error control, we will show the expectation that ϱ tends to α, and its variance tends to zero. For the expectation, for

any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}[\varrho] &= \frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(\left|v^\top t_j\right| > z_{\frac{\alpha}{2}}\right) \\
&\leq \frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} \left[\mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(|\varsigma_j| > \epsilon\right)\right] \\
&= \frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(|\varsigma_j| > \epsilon\right) \\
&\leq \frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(\max_{1 \leq j \leq p} |\varsigma_j| > \epsilon\right) \to 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right),
\end{aligned}$$

where the last convergence holds because the Cesaro mean converges to the same limit as

$$\lim_{n,p} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon\right) = 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right),$$

while the term $\mathbb{P}(\max_{1 \leq j \leq p} |\varsigma_j| > \epsilon)$ varnishes. Similarly, we can show that $\liminf_{n,p \to \infty} \mathbb{E}[\varrho] \geq 2\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right)$ for all $\epsilon > 0$. Let $\epsilon \to 0^+$, it follows that $\mathbb{E}[\varrho] \to \alpha$ as $n, p \to \infty$.

For any $\epsilon > 0$, the second moment can be upped bounded as below:

$$\begin{aligned}
\mathbb{E}[\varrho^2] &= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k \in \mathcal{N}_p} \mathbb{P}\left(\left|v^\top t_j\right| > z_{\frac{\alpha}{2}}, \left|v^\top t_k\right| > z_{\frac{\alpha}{2}}\right) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(\left|v^\top t_j\right| > z_{\frac{\alpha}{2}}\right) + \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k \in \mathcal{N}_p, j \neq k} \mathbb{P}\left(\left|v^\top t_j\right| > z_{\frac{\alpha}{2}}, \left|v^\top t_k\right| > z_{\frac{\alpha}{2}}\right) \\
&\leq \frac{1}{|\mathcal{N}_p|^2} \sum_{j \in \mathcal{N}_p} \mathbb{P}\left(\left|v^\top t_j\right| > z_{\frac{\alpha}{2}}\right) \\
&\quad + \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k \in \mathcal{N}_p, j \neq k} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon, |\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon\right) + \mathbb{P}\left(|\varsigma_j| > \epsilon\right) + \mathbb{P}\left(|\varsigma_k| > \epsilon\right) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k \in \mathcal{N}_p, j \neq k} \mathbb{P}\left(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon, |\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon\right) + o(1) \\
&= \frac{1}{|\mathcal{N}_p|^2} \sum_{j,k \in \mathcal{N}_p, j \neq k} \mathbb{P}(|\vartheta_j| > z_{\frac{\alpha}{2}} - \epsilon \mid X, \widehat{U}) \mathbb{P}(|\vartheta_k| > z_{\frac{\alpha}{2}} - \epsilon \mid X, \widehat{U}) + o(1) \\
&\to 4\left(1 - \Phi\left(z_{\frac{\alpha}{2}} - \epsilon\right)\right)^2,
\end{aligned}$$

where the last equality is from the independence of ϑ_j and ϑ_k. We can similarly obtain the lower bound. Let $\epsilon \to 0^+$, it follows that $\mathbb{E}[\varrho^2] \to \alpha^2$ and $\mathbb{V}(\varrho) \to 0$ as $n, p \to \infty$. Combining the previous results yields that $\varrho \xrightarrow{\mathrm{p}} \alpha$. □

C.5.3 Nonlinear modeling

The natural extension of partial linear models to the nonlinear cases is the generalized partially linear models [65, 154]:

$$g(\mathbb{E}[Y \mid X, U]) = \beta^\top X + h(U), \tag{C.5.1}$$

by introducing a proper link function g, applied element-wisely on the conditional mean of the outcomes. Similar to the results in the previous sections, a nonlinear counterpart of the main effect estimand (4.2.5) is given by

$$\beta(\mathbb{P}) = \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}\mathbb{E}[\mathrm{Cov}[X, g(\mathbb{E}(Y|X,U)) \mid U]]. \tag{C.5.2}$$

Such an estimand has been considered in Newey and Robins [130], Robins et al. [141] with the identity link and in Vansteelandt and Dukes [171] with a single treatment. When the model (C.5.1) is correctly specified, (C.5.2) is equivalent to the regression coefficient under model (C.5.1). On the other hand, when the model (C.5.1) is misspecified, estimand (C.5.2) still represents a meaningful statistical quantity.

With a differentiable link function g, the influence function (for $\widetilde{\Sigma}\widetilde{\beta}$) analoguous to (4.3.3) is given by:

$$\widetilde{\varphi}(O;\mathbb{P}) := (X - \mathbb{E}[X \mid \widehat{U}])(\eta(O) - \widetilde{\beta}^\top(X - \mathbb{E}[X \mid \widehat{U}]))^\top,$$

where the main effect estimand with estimated embedding is defined as:

$$\widetilde{\beta} = \mathbb{E}[\mathrm{Cov}(X \mid \widehat{U})]^{-1}\mathbb{E}[\mathrm{Cov}[X, g(\mathbb{E}(Y|X,\widehat{U})) \mid \widehat{U}]], \tag{C.5.3}$$

and the function η is defined as:

$$\eta(O) = g'(\mathbb{E}[Y \mid X, \widehat{U}]) \odot (Y - \mathbb{E}[Y \mid X, \widehat{U}]) + g(\mathbb{E}[Y \mid X, \widehat{U}]) - \mathbb{E}[g(\mathbb{E}[Y \mid X, \widehat{U}]) \mid \widehat{U}].$$

The doubly robust semiparametric inference results in Theorem 24 and Corollary 25 can be extended to accommodate nonlinear link functions, as shown in the next theorem.

Theorem C.5.1 (Doubly robust inference with nonlinear link functions)**.** Under a nonparameteric model and a differentiable link function g, define the estimator of β in (C.5.2) as:

$$\widehat{\beta} = \mathbb{P}_n\{(X - \widehat{\mathbb{E}}(X \mid \widehat{U}))^{\otimes 2}\}^{-1}\mathbb{P}_n\{(X - \widehat{\mathbb{E}}(X \mid \widehat{U})) \cdot (\mathbb{I} - \mathbb{P}_n)\{g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}])\}^\top\}, \tag{C.5.4}$$

which depends on empirical measure $\mathbb{P}_n$ and two nuisance functions $\widehat{\mathbb{E}}[X \mid \widehat{U}]$ and $\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]$ estimated from independent samples of $\mathbb{P}_n$. Under Assumptions 12 and 14 and assume that

1. Local Lipschitzness: There exists $L > 0$ such that $\|g(\mathbb{E}[Y \mid X, \widehat{U}]) - g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) - g'(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) \odot (\mathbb{E}[Y \mid X, \widehat{U}] - \widehat{\mathbb{E}}[Y \mid X, \widehat{U}])\|_\infty \leq L\|\mathbb{E}[Y \mid X, \widehat{U}] - \widehat{\mathbb{E}}[Y \mid X, \widehat{U}]\|_\infty^2$.
2. Boundeness and consistency: Assumptions 12 and 14 hold with additionally, $\|\eta(O)\|_{\mathrm{L}_{2(1+\delta^{-1})}} < M$ and $\|\|\widehat{\eta}(O) - \eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} = o_{\mathbb{P}}(1)$.
3. Rate condition: $\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2}^2$, $\|\mathbb{E}[Y \mid X, \widehat{U}] - \widehat{\mathbb{E}}[Y \mid X, \widehat{U}]\|_{\mathrm{L}_2,\infty}^2$, and $\|\mathbb{E}[g(\mathbb{E}[Y \mid X, \widehat{U}]) \mid \widehat{U}] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) \mid \widehat{U}]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X \mid \widehat{U}] - \widehat{\mathbb{E}}[X \mid \widehat{U}]\|_{\mathrm{L}_2}$ are of order $o_{\mathbb{P}}(n^{-\frac{1}{2}})$.

Then, the estimator $\widetilde{b}$ is asymptotically normal:

$$\sqrt{n}(\widetilde{b}_{\cdot j} - \widetilde{\beta}_{\cdot j}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, \widetilde{\Sigma}^{-1}\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O;\mathbb{P})\}\widetilde{\Sigma}^{-1}), \qquad j = 1, \ldots, p.$$

Furthermore, if the conditions of Theorem 22 hold with $\ell_m = o(n^{-\frac{1}{2}})$, then we have

$$\sqrt{n}(\widetilde{b}_{\cdot j} - \beta_{\cdot j}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, \widetilde{\Sigma}^{-1}\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O;\mathbb{P})\}\widetilde{\Sigma}^{-1}), \qquad j = 1, \ldots, p.$$

Algorithm C.5.7 Semiparametric inference for main effects with nonlinear link functions

Input: Reponses Y, covariate X, estimated latent embedding $\widehat{U}$, and link function g.

1: Use machine learning methods to obtain nuisance estimates $\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]$ and $\widehat{\mathbb{E}}[X \mid \widehat{U}]$.

2: Use a data-adaptive fit $g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) \sim \widehat{U}$ to obtain estimated regression function $\widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) \mid \widehat{U}]$. If X is categorical with finite support $|\mathcal{X}| < \infty$, this simply reduces to $\widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y \mid X, \widehat{U}]) \mid \widehat{U}] = \sum_{x \in \mathcal{X}} g(\widehat{\mathbb{E}}[Y \mid X = x, \widehat{U}])\widehat{\mathbb{E}}[X = x \mid \widehat{U}]$.

3: Fit a linear regression of $\widehat{\eta}(O) \sim X - \widehat{\mathbb{E}}[X \mid \widehat{U}]$ without an intercept to obtain an estimate $\widetilde{b}$ as defined in (C.5.4) of $\widetilde{\beta}$ as defined in (C.5.3).

4: Estimate the variance of $\widetilde{b}_{\cdot j}$ by $\widehat{S}_j/n$ based on Theorem C.5.1, where $\widehat{S}_j = \widehat{\Sigma}^{-1}\mathbb{V}_n\{\widetilde{\varphi}_{\cdot j}(O; \widehat{\mathbb{P}})\}\widehat{\Sigma}^{-1}$.

Output: Confidence intervals and p-values based on asymptotic null distribution $\widetilde{b}_{\cdot j} \dot{\sim} \mathcal{N}_d(\widetilde{\beta}_{\cdot j}, \frac{\widehat{S}_j}{n})$.

Compared to Theorem 24, Theorem C.5.1 requires additional assumptions regarding the Lipschitzness of the link function around the true regression function, as noted by Vansteelandt and Dukes [171]. It also requires boundedness and consistency assumptions on the first-order expansion term η. Nevertheless, the overall conclusion is similar when both the estimators and the influence functions have a different link function for a different target estimand. The double robustness still allows efficient semiparametric inference with data-adaptive estimation procedures.

Proof of Theorem C.5.1. From Lemma C.5.3, we have

$$\sqrt{n}(\widetilde{b} - \widetilde{\beta}) = \sqrt{n}\mathbb{P}\{(X - \mathbb{E}(X \mid \widehat{U}))^{\otimes 2}\}^{-1}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(O; \mathbb{P})\} + \widetilde{\xi}$$

where $\widetilde{\varphi}$ is defined as

$$\widetilde{\varphi}(O; \mathbb{P}) = (X - \mathbb{E}[X \mid \widehat{U}])(\eta(O) - \widetilde{\beta}^\top(X - \mathbb{E}[X \mid \widehat{U}]))^\top \tag{C.5.5}$$

and the remainder term $\widetilde{\xi}$ satisfies that $\|\widetilde{\xi}\|_{2,\infty} = o_{\mathbb{P}}(1)$. This proves the first statement.

When $\|\widehat{U} - U\|_{\mathrm{L}_2} = o_{\mathbb{P}}(n^{-\frac{1}{2}})$, from Theorem 22 we have $\|\widetilde{\beta} - \beta\|_{2,\infty} = o_{\mathbb{P}}(n^{-\frac{1}{2}})$. Therefore, we further have

$$\sqrt{n}(\widetilde{b} - \beta) = \sqrt{n}(\widetilde{b} - \widetilde{\beta}) + \sqrt{n}(\widetilde{\beta} - \beta) = \sqrt{n}\mathbb{P}\{(X - \mathbb{E}(X \mid \widehat{U}))^{\otimes 2}\}^{-1}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(O; \mathbb{P})\} + \xi,$$

with $\|\xi\|_{2,\infty} = o_{\mathbb{P}}(1)$.

To establish the asymptotic normality, we apply the triangle-array CLT in Lemma C.5.4. This requires verifying the sufficient condition of the Lindeberg condition. Because $\mathbb{V}\{\widetilde{\Sigma}^{-1}\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\} = \widetilde{\Sigma}^{-1}\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}\widetilde{\Sigma}^{-1}$, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{V}\{\widetilde{\Sigma}^{-1}\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{-\frac{1}{2}}(\widetilde{\Sigma}^{-1}\widetilde{\varphi}_{\cdot j}(O; \mathbb{P}))\|^{2+\frac{2}{\delta}}] \\
=&\mathbb{E}[\|\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{-\frac{1}{2}}\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\|^{2+\frac{2}{\delta}}] \\
\leq&\|\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{-\frac{1}{2}}\|_{\mathrm{op}}^{2+\frac{2}{\delta}} \cdot \mathbb{E}[\|\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\|^{2+\frac{2}{\delta}}] \\
\leq&\|\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{-\frac{1}{2}}\|_{\mathrm{op}}^{2+\frac{2}{\delta}} \cdot (\mathbb{E}[\|(X - \mathbb{E}[X \mid \widehat{U}])\eta_j(O)\|^{2+\frac{2}{\delta}}] + \mathbb{E}[\|(X - \mathbb{E}[X \mid \widehat{U}])^{\otimes 2}\widetilde{\beta}_{\cdot j}\|^{2+\frac{2}{\delta}}]) \\
\leq&\|\mathbb{V}\{\widetilde{\varphi}_{\cdot j}(O; \mathbb{P})\}^{-\frac{1}{2}}\|_{\mathrm{op}}^{2+\frac{2}{\delta}} \cdot \mathbb{E}[\|X - \mathbb{E}[X \mid \widehat{U}]\|^{1+\frac{1}{\delta}}\|\eta(O)\|^{1+\frac{1}{\delta}}] + \|\widetilde{\beta}_{\cdot j}\|^{2+\frac{2}{\delta}} \\
\leq&\sigma^{-1-\frac{1}{\delta}}M^{2+\frac{2}{\delta}} + \|\widetilde{\beta}_{\cdot j}\|^{2+\frac{2}{\delta}}.
\end{aligned}$$

Now applying Lemma C.5.4 finishes the proof. □

C.5.4 Auxillary lemmas

Lemma C.5.2 (Efficient influence function)**.** Consider a random variable $O = (X, U, Y) \in \mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p$ under a nonparameteric model and a differentiable function g, the main effect estimand in $\mathbb{R}^{d\times p}$:

$$\beta = \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}\mathbb{E}[\mathrm{Cov}(X, g(\mathbb{E}[Y \mid X, U]) \mid U)],$$

(where g is applied entry-wisely) has an efficient influence function $\mu : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^{d\times p}$ given by:

$$\varphi(O) = \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}(X - \mathbb{E}[X \mid U])(\eta(O) - \beta^\top(X - \mathbb{E}[X \mid U]))^\top,$$

where $\eta : \mathbb{R}^r \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^p$ is defined as:

$$\eta(O) = g'(\mathbb{E}[Y \mid X, U]) \odot (Y - \mathbb{E}[Y \mid X, U]) + g(\mathbb{E}[Y \mid X, U]) - \mathbb{E}[g(\mathbb{E}[Y \mid X, U]) \mid U].$$

Proof of Lemma C.5.2. The proof follows similarly as in Vansteelandt and Dukes [171, Theorem 1] for a univariate treatment and a univariate outcome, and extends the previous results to the multivariate cases. Below, we present a simplified derivation of the influence function.

Under the nonparametric model for the observed data $O = (X, U, Y)$. We first calculate the efficient influence function of

$$\begin{aligned}\theta(\beta) &= \mathbb{E}[(X - \mathbb{E}[X \mid U]))(g(\mathbb{E}[Y \mid X, U]) - \beta^\top(X - \mathbb{E}[X \mid U]))^\top] \\ &= \int (X - \mathbb{E}[X \mid U]))(g(\mathbb{E}[Y \mid X, U]) - \beta^\top X)\,\mathrm{d}P(O),\end{aligned}$$

where $P(O)$ is the joint distribution of data. Note that by the definition of β, we have $\theta(\beta) = 0$.

Consider a one-dimensional submodel of $p(O)$ indexed by a scalar parameter t, and let $S_t(o) = \partial \log \mathrm{d}P_t(o)/\partial t \mid_{t=0}$ denote the score function of the submodel. Similarly, let $S_t(Y \mid X, U)$, $S_t(X \mid U)$ and $S_t(U)$ be the scores w.r.t. t in that parametric submodel, corresponding to the distributions $p(Y \mid X, U)$, $p(X \mid U)$ and $p(U)$, respectively Taking the derivative of θ w.r.t. t, we obtain

$$\begin{aligned}\frac{\partial\theta(\beta)}{\partial t}\Big|_{t=0} &= \int \frac{\partial(X - \mathbb{E}_t[X \mid U])}{\partial t}\Big|_{t=0}(g(\mathbb{E}[Y \mid X, U]) - \beta^\top X)^\top\,\mathrm{d}P(O) \\ &\quad + \int (X - \mathbb{E}[X \mid U]))\left(g'(\mathbb{E}[Y \mid X, U]) \odot \frac{\partial\mathbb{E}[Y \mid X, U]}{\partial t}\Big|_{t=0}\right)^\top \mathrm{d}P(O) \\ &\quad + \int (X - \mathbb{E}[X \mid U]))(g(\mathbb{E}[Y \mid X, U]) - \beta^\top X)^\top \frac{\partial p_t(X, U)}{\partial t}\Big|_{t=0}\mathrm{d}O \\ &= -\int (X - \mathbb{E}[X \mid U])\mathbb{E}[g(\mathbb{E}[Y \mid X, U]) - \beta^\top X \mid U]^\top S_t(X \mid U)\,\mathrm{d}P(O) \\ &\quad + \int (X - \mathbb{E}[X \mid U]))(g'(\mathbb{E}[Y \mid X, U]) \odot (Y - \mathbb{E}[Y \mid X, U]))^\top S_t(Y \mid X, U)\,\mathrm{d}P(O) \\ &\quad + \int (X - \mathbb{E}[X \mid U]))(g(\mathbb{E}[Y \mid X, U]) - \beta^\top X)^\top S_t(X, U)\,\mathrm{d}P(O),\end{aligned}$$

where in the first equality, we apply the product and chain rules [84, Section 3.4.3]; and in the second equality, we use the identity $S_t(Z) = \partial \log p_t(Z)/\partial t = (\partial p_t(Z)/\partial t)/p_t(Z)$ for score functions.

Note that

$$S_t(O) = S_t(Y \mid X, U) + S_t(X \mid U) + S_t(U).$$

From the zero mean properties of scores and $\theta(\beta) = 0$, we further have

$$\begin{aligned}
\frac{\partial \theta(\beta)}{\partial t}\Big|_{t=0} &= -\int (X - \mathbb{E}[X \mid U])\mathbb{E}[g(\mathbb{E}[Y \mid X, U]) - \beta^\top X \mid U]^\top S_t(O)\,\mathrm{d}P(O) \\
&\quad + \int (X - \mathbb{E}[X \mid U]))(g'(\mathbb{E}[Y \mid X, U]) \odot (Y - \mathbb{E}[Y \mid X, U]))^\top S_t(O)\,\mathrm{d}P(O) \\
&\quad + \int (X - \mathbb{E}[X \mid U]))(g(\mathbb{E}[Y \mid X, U]) - \beta^\top X)^\top S_t(O)\,\mathrm{d}P(O) \\
&= \int (X - \mathbb{E}[X \mid U]))(\eta(O) - \beta^\top (X - \mathbb{E}[X \mid U]))^\top S_t(O)\,\mathrm{d}P(O),
\end{aligned}$$

which implies that $(X - \mathbb{E}[X \mid U]))(\eta(O) - \beta^\top (X - \mathbb{E}[X \mid U]))^\top$ is an influence function for θ. From a similar argument in the proof of Theorem 1 in Vansteelandt and Dukes [171], it is also the efficient influence function of $\theta(\beta)$ under the nonparametric model. Consequently, by chain rule $\partial\theta/\partial t = (\partial\theta/\partial\beta)(\partial\beta/\partial t)$, the conclusion follows by taking the inverse of $\partial\theta/\partial\beta$. □

Remark 15 (Alternative expression of the estimand)**.** Note that the first part of the influence function also gives an alternative expression for β:

$$\beta = \mathbb{E}[\mathrm{Cov}(X \mid U)]^{-1}\mathbb{E}[(X - \mathbb{E}[X \mid U])\eta(O)^\top] \tag{C.5.6}$$

because

$$\mathbb{E}[(X - \mathbb{E}[X \mid U])(g'(\mathbb{E}[Y \mid X, U]) \odot (Y - \mathbb{E}[Y \mid X, U]))^\top] = 0, \tag{C.5.7}$$

by the law of iterated expectation.

Lemma C.5.3 (Doubly robust estimation)**.** Consider the setting in Lemma C.5.2. Define a plug-in estimator of β:

$$\widehat{\beta} = \mathbb{P}_n\{(X - \widehat{\mathbb{E}}(X \mid U))^2\}^{-1}\mathbb{P}_n\{(X - \widehat{\mathbb{E}}(X \mid U)) \cdot (\mathbb{I} - \mathbb{P}_n)\{g(\widehat{\mathbb{E}}[Y \mid X, U])\}^\top\}$$

which depends on empirical measure $\mathbb{P}_n$ and two nuisance functions $\widehat{\mathbb{E}}[X \mid U]$ and $\widehat{\mathbb{E}}[Y \mid X, U]$ estimated from independent samples of $\mathbb{P}_n$. Define the population and empirical variance by

$$\begin{aligned}
\Sigma &:= \mathbb{P}\{(X - \mathbb{E}(X \mid U))^{\otimes 2}\} \\
\widehat{\Sigma} &:= \mathbb{P}_n\{(X - \widehat{\mathbb{E}}(X \mid U))^{\otimes 2}\},
\end{aligned}$$

the empirical influence function (for $\Sigma\beta$) by:

$$\varphi(O; \widehat{\mathbb{P}}) := (X - \widehat{\mathbb{E}}[X \mid U])(\widehat{\eta}(O) - \widehat{\beta}^\top (X - \widehat{\mathbb{E}}[X \mid U]))^\top.$$

Suppose the following conditions hold:

- (Regularity conditions) There exists $\sigma > 0$ such that $\Sigma \succeq \sigma I_d$, $\widehat{\Sigma} \succeq \sigma I_d$.
- (Bounded moments and consistency) There exists $\delta \in (0, 1]$ and $M > 0$, such that

$$\|\beta\|_{2,\infty} \vee \|X - \mathbb{E}[X \mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}} \vee \|X - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}} \vee \|\eta(O)\|_{\mathrm{L}_{2(1+\delta^{-1})}} < M$$

$$\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}}, \|\|\widehat{\eta}(O) - \eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} = o_{\mathbb{P}}(1)$$

- (Local Lipshitzness) There exists $L > 0$ such that

$$\begin{aligned}&\|g(\mathbb{E}[Y \mid X, U]) - g(\widehat{\mathbb{E}}[Y \mid X, U]) - g'(\widehat{\mathbb{E}}[Y \mid X, U]) \odot (\mathbb{E}[Y \mid X, U] - \widehat{\mathbb{E}}[Y \mid X, U])\|_\infty \\ \leq & L\|\mathbb{E}[Y \mid X, U] - \widehat{\mathbb{E}}[Y \mid X, U])\|_\infty^2. \end{aligned} \tag{C.5.8}$$

Then, it holds that

$$\sqrt{n}(\widehat{\beta} - \beta) = \sqrt{n}\Sigma^{-1}(\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + \xi,$$

where for any $\epsilon > 0$, there exists a constant $C = C(\epsilon, \sigma, M, L)$, such that with probability at least $1 - 3\epsilon$, the remainder term satisfies that

$$\begin{aligned}\|\xi\|_{2,\infty} \leq & C\{\|(\mathbb{P}_n - \mathbb{P})\{(X - \mathbb{E}[X \mid U])^{\otimes 2}\}\|_{\text{op}} + \|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}} + \|\|\eta(O) - \widehat{\eta}(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}}\} \\ & + C\sqrt{n}\{\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_2}^2 \\ & \quad + ML\|\mathbb{E}[Y \mid X, U] - \widehat{\mathbb{E}}[Y \mid X, U]\|_{\mathrm{L}_2,\infty}^2 \\ & \quad + \|\mathbb{E}[g(\mathbb{E}[Y \mid X, U]) \mid U] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y \mid X, U]) \mid U]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_2}\},\end{aligned}$$

When $\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_2}^2$, $\|\mathbb{E}[Y \mid X, U] - \widehat{\mathbb{E}}[Y \mid X, U]\|_{\mathrm{L}_2,\infty}^2$, and $\|\mathbb{E}[g(\mathbb{E}[Y \mid X, U]) \mid U] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y \mid X, U]) \mid U]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_2}$ are of order $o_{\mathbb{P}}(n^{-\frac{1}{2}})$, we further have that $\|\xi\|_{2,\infty} = o_{\mathbb{P}}(1)$ and hence

$$\sqrt{n}(\widehat{\beta}_{\cdot j} - \beta_{\cdot j}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, \Sigma^{-1}\mathbb{V}\{\varphi_{\cdot j}(O; \mathbb{P})\}), \qquad j = 1, \ldots, p.$$

Proof of Lemma C.5.3. From the definition of $\widehat{\beta}$, we have $\mathbb{P}_n\{\varphi(O; \widehat{\mathbb{P}})\} = 0$. Therefore, $\widehat{\beta}$ is also a one-step estimator. We begin with a three-term decomposition of the estimation error (see, for example, Du et al. [49, Equation (2.2)] and Kennedy [84, Equation (10)]):

$$\begin{aligned}\widehat{\Sigma}\sqrt{n}(\widehat{\beta} - \beta) = & \sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} \\ & + \sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\varphi(O; \widehat{\mathbb{P}}) - \varphi(O; \mathbb{P})\} + \sqrt{n}(\widehat{\Sigma} - \widetilde{\Sigma})(\widehat{\beta} - \beta) \\ & + \sqrt{n}\widetilde{\Sigma}(\widehat{\beta} - \beta) + \sqrt{n}\mathbb{P}\{\varphi(O; \widehat{\mathbb{P}})\} \\ =: & C + T_1 + T_2,\end{aligned} \tag{C.5.9}$$

where $\widetilde{\Sigma} := \mathbb{P}\{(X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2}\}$. By the central limit theorem, each entry of the first term C is $\mathcal{O}_{\mathbb{P}}(1)$. We next derive finite-sample deviation bounds for the other terms and show that they are $o_{\mathbb{P}}(1)$ under the extra rate conditions as assumed.

Part (1) Controlling the empirical process term T_1. We begin by decomposing T_1:

$$\begin{aligned}& \varphi(O; \widehat{\mathbb{P}}) - \varphi(O; \mathbb{P}) + (X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2}(\widehat{\beta} - \beta) \\ = & (X - \widehat{\mathbb{E}}[X \mid U])(\widehat{\eta}(O) - \widehat{\beta}^\top(X - \widehat{\mathbb{E}}[X \mid U]))^\top - (X - \mathbb{E}[X \mid U])(\eta(O) - \beta^\top(X - \mathbb{E}[X \mid U]))^\top \\ & \quad + (X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2}(\widehat{\beta} - \beta) \\ = & [(X - \mathbb{E}[X \mid U])^{\otimes 2} - (X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2}]\beta + [(X - \widehat{\mathbb{E}}[X \mid U])\widehat{\eta}(O)^\top - (X - \mathbb{E}[X \mid U])\eta(O)^\top] \\ = & : S_1 + S_2.\end{aligned}$$

Note that each term above takes the form of $\widehat{a}\widehat{b} - ab = \widehat{a}(\widehat{b} - b) + (\widehat{a} - a)b$, which we will next use to derive the upper bound.

For the first term, we have

$$\begin{aligned}
&\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})S_1\|_{2,\infty} \\
=&\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})[(X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2} - (X - \mathbb{E}[X \mid U])^{\otimes 2}]\beta\|_{2,\infty} \\
=&\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A_1\}\beta\|_{2,\infty},
\end{aligned}$$

where

$$A_1 = (X - \widehat{\mathbb{E}}[X \mid U])^{\otimes 2} - (X - \mathbb{E}[X \mid U])^{\otimes 2}.$$

From Lemma C.5.5, we have

$$\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})S_1\|_{2,\infty} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A_1\|_{\text{op}}^2\|\beta\|_{2,\infty}^2]^{\frac{1}{2}} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A_1\|_{\text{op}}^2]^{\frac{1}{2}}\|\beta\|_{2,\infty},$$

with probability at least $1-\epsilon$. Now, it remains to derive the upper bound of the expected squared operator norm:

$$\begin{aligned}
\mathbb{E}[\|A_1\|_{\text{op}}^2]^{\frac{1}{2}} &\leq \mathbb{E}[\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_2^2(\|X - \mathbb{E}[X \mid U]\|_2 + \|X - \widehat{\mathbb{E}}[X \mid U]\|_2)^2]^{\frac{1}{2}} \\
&\leq \|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}}(\|X - \mathbb{E}[X \mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}} + \|X - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}}).
\end{aligned}$$

Therefore, we have

$$\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})S_1\|_{2,\infty} \leq 2M^2\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}}$$

with probability at least $1 - \epsilon$.

For the second term, similarly, we have

$$\begin{aligned}
&\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})S_2\|_{2,\infty} \\
\leq&\epsilon^{-\frac{1}{2}}\max_{j\in[p]}\mathbb{E}[\|(X - \widehat{\mathbb{E}}[X \mid U])[(\widehat{\eta}(O) - \eta(O))^\top]_{\cdot j} + (\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U])[\eta(O)^\top]_{\cdot j}\|^2]^{\frac{1}{2}} \\
\leq&\epsilon^{-\frac{1}{2}}(\|X - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}}\|\|(\widehat{\eta}(O) - \eta(O))^\top\|_{2,\infty}^2\|_{\mathrm{L}_{1+\delta}}^{\frac{1}{2}} + \|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}}\|\eta(O)^\top\|_{\mathrm{L}_{2(1+\delta^{-1})},\infty}) \\
\leq&\epsilon^{-\frac{1}{2}}M(\|\|\widehat{\eta}(O) - \eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} + \|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}})
\end{aligned}$$

with probability at least $1 - \epsilon$.

Combining the above results, with probability at least $1 - 2\epsilon$, we have

$$\begin{aligned}
\|T_1\|_{2,\infty} &\leq 2M^2\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}} \\
&\quad + M(\|\|\widehat{\eta}(O) - \eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}} + \|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}}) \\
&\leq 2M(M \vee 1)\|\mathbb{E}[X \mid U] - \widehat{\mathbb{E}}[X \mid U]\|_{\mathrm{L}_{2(1+\delta)}} + M\|\|\widehat{\eta}(O) - \eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}}. \qquad \text{(C.5.10)}
\end{aligned}$$

Part (2) Controlling the bias term T_2. For the third term T_2 in (C.5.9), we have

$$\begin{aligned}
T_2 &= \sqrt{n}\widetilde{\Sigma}(\widehat{\beta}-\beta) + \sqrt{n}\mathbb{P}\{\varphi(O;\widehat{\mathbb{P}})\} \\
&= \sqrt{n}\mathbb{P}\{(X-\widehat{\mathbb{E}}[X\mid U])\widehat{\eta}(O)^\top\} - \sqrt{n}\widetilde{\Sigma}\beta \\
&= \sqrt{n}\mathbb{P}\{(X-\widehat{\mathbb{E}}[X\mid U])(g'(\widehat{\mathbb{E}}[Y\mid X,U])\odot(Y-\widehat{\mathbb{E}}[Y\mid X,U]) + g(\widehat{\mathbb{E}}[Y\mid X,U]) - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U])^\top\} \\
&\quad - \sqrt{n}\widetilde{\Sigma}\Sigma^{-1}\Sigma\beta, \qquad \text{(C.5.11)}
\end{aligned}$$

where the last equality is because of Equations (C.5.6) and (C.5.7). Denote the second-order remaining term by $Q = g(\mathbb{E}[Y\mid X,U]) - g(\widehat{\mathbb{E}}[Y\mid X,U]) - g'(\widehat{\mathbb{E}}[Y\mid X,U])\odot(\mathbb{E}[Y\mid X,U] - \widehat{\mathbb{E}}[Y\mid X,U])$. Then, we further have

$$\begin{aligned}
T_2 &= -\sqrt{n}\mathbb{P}\{(X-\widehat{\mathbb{E}}[X\mid U])(g(\mathbb{E}[Y\mid X,U]) + \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U] + Q)^\top\} \\
&\quad + \sqrt{n}\mathbb{P}\{(X-\mathbb{E}[X\mid U])(g(\mathbb{E}[Y\mid X,U]) - \mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U])^\top\} \\
&\quad + \sqrt{n}(I_d - \widetilde{\Sigma}\Sigma^{-1})\Sigma\beta \\
&= \sqrt{n}\mathbb{P}\{(X-\widehat{\mathbb{E}}[X\mid U])Q^\top\} \\
&\quad + \sqrt{n}\mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])(\mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U])^\top\} \\
&\quad + \sqrt{n}(\Sigma - \widetilde{\Sigma})\beta. \qquad \text{(C.5.12)}
\end{aligned}$$

Because by the law of iterative expectation,

$$\mathbb{P}\{(X-\mathbb{E}[X\mid U])(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^\top\} = 0, \qquad \text{(C.5.13)}$$

we have

$$\begin{aligned}
1 - \widetilde{\Sigma}\Sigma^{-1} &= 1 - \mathbb{P}\{(X-\mathbb{E}[X\mid U] + \mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\}\Sigma^{-1} \\
&= -\mathbb{P}\{(X-\mathbb{E}[X\mid U])(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^\top\} - \mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])(X-\mathbb{E}[X\mid U])^\top\} \\
&\quad + \mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\}\Sigma^{-1} \\
&= \mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\}\Sigma^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\|\Sigma - \widetilde{\Sigma}\|_{\text{op}} &= \|(I_d - \widetilde{\Sigma}\Sigma^{-1})\Sigma\|_{\text{op}} \qquad \text{(C.5.14)} \\
&= \|\mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\}\|_{\text{op}} \\
&\le \|\mathbb{P}\{(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\}\|_{\text{op}} \qquad \text{(Jensen's inequality)} \\
&\le \mathbb{P}\{\|(\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U])^{\otimes 2}\|_{\text{op}}\} \\
&= \|\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U]\|_{\text{L}_2}^2 \qquad \text{(C.5.15)}
\end{aligned}$$

Combining Equations (C.5.8), (C.5.12) and (C.5.15) yields that

$$\begin{aligned}
\|T_2\|_{2,\infty} &\le \sqrt{n}\|Q\|_{\text{L}_2,\infty}\|X - \widehat{\mathbb{E}}[X\mid U]\|_{\text{L}_2} \\
&\quad + \sqrt{n}\|\mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U]\|_{\text{L}_2,\infty}\|\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U]\|_{\text{L}_2} \\
&\quad + \sqrt{n}\|\Sigma - \widetilde{\Sigma}\|_{\text{op}}\|\beta\|_{2,\infty} \\
&\le \sqrt{n}ML\|\mathbb{E}[Y\mid X,U] - \widehat{\mathbb{E}}[Y\mid X,U]\|_{\text{L}_2,\infty}^2 \\
&\quad + \sqrt{n}\|\mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U] - \widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U]\|_{\text{L}_2,\infty}\|\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U]\|_{\text{L}_2} \\
&\quad + M\sqrt{n}\|\mathbb{E}[X\mid U] - \widehat{\mathbb{E}}[X\mid U]\|_{\text{L}_2}^2. \qquad \text{(C.5.16)}
\end{aligned}$$

Part (3) Combining the above results. Finally, from Equations (C.5.9), (C.5.10) and (C.5.16)

$$\widehat{\Sigma}\sqrt{n}(\widehat{\beta}-\beta)=\sqrt{n}(\mathbb{P}_n-\mathbb{P})\{\varphi(O;\mathbb{P})\}+\varsigma$$

for some $\varsigma\in\mathbb{R}^{d\times p}$ with

$$\begin{aligned}\|\varsigma\|_{2,\infty}\le{}& 2M(M\vee 1)\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_{2(1+\delta)}}+\sqrt{n}M\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}^2\\&+M\|\|\widehat{\eta}(O)-\eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}}\\&+\sqrt{n}ML\|\mathbb{E}[Y\mid X,U]-\widehat{\mathbb{E}}[Y\mid X,U]\|_{\mathrm{L}_2,\infty}^2\\&+\sqrt{n}\|\mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U]-\widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}.\end{aligned}$$

Note that

$$\begin{aligned}\|\Sigma^{-1}-\widehat{\Sigma}^{-1}\|_{\mathrm{op}}&=\|\widehat{\Sigma}^{-1}(\widehat{\Sigma}-\Sigma)\Sigma^{-1}\|_{\mathrm{op}}\\&\le\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}\|\widehat{\Sigma}-\Sigma\|\|_{\mathrm{op}}\|\Sigma^{-1}\|_{\mathrm{op}}\\&\le\sigma^2\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}\\&\le\sigma^2\|(\mathbb{P}_n-\mathbb{P})\{(X-\mathbb{E}[X\mid U])^{\otimes 2}\}\|_{\mathrm{op}}+\sigma^2\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}^2,\end{aligned}$$

where the first equality is from $\widehat{\Sigma}^{-1}(\widehat{\Sigma}-\Sigma)\Sigma^{-1}=\Sigma^{-1}-\widehat{\Sigma}^{-1}$, the second inequality is from the positivity assumption that $\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}\le\sigma$, $\|\Sigma^{-1}\|_{\mathrm{op}}\le\sigma$, and the last inequality is from (C.5.15). We further have

$$\sqrt{n}(\widehat{\beta}-\beta)=\sqrt{n}\Sigma^{-1}(\mathbb{P}_n-\mathbb{P})\{\varphi(O;\mathbb{P})\}+\xi,$$

with

$$\xi=\sqrt{n}(\widehat{\Sigma}^{-1}-\Sigma^{-1})(\mathbb{P}_n-\mathbb{P})\{\varphi(O;\mathbb{P})\}+\widehat{\Sigma}^{-1}\varsigma.$$

By multidimensional Chebyshev inequality and union bound, with probability at least $1-3\epsilon$,

$$\begin{aligned}\|\xi\|_{2,\infty}&\le\sigma^2(\|(\mathbb{P}_n-\mathbb{P})\{(X-\mathbb{E}[X\mid U])^{\otimes 2}\}\|_{\mathrm{op}}+\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}^2)\\&\quad\cdot\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_{2(1+\delta^{-1})}}(\|\eta(O)\|_{\mathrm{L}_{2(1+\delta)},\infty}+\|\beta\|_{2,\infty})+\sigma\|\varsigma\|_{2,\infty}\\&\le 2\sigma^2M^2\|(\mathbb{P}_n-\mathbb{P})\{(X-\mathbb{E}[X\mid U])^{\otimes 2}\}\|_{\mathrm{op}}\\&\quad+2M(M\vee 1)\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_{2(1+\delta)}}+M\|\|\widehat{\eta}(O)-\eta(O)\|_\infty\|_{\mathrm{L}_{2(1+\delta)}}\\&\quad+\sqrt{n}2(\sigma^2\vee 1)M(M\vee 1)\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}^2\\&\quad+\sqrt{n}ML\|\mathbb{E}[Y\mid X,U]-\widehat{\mathbb{E}}[Y\mid X,U]\|_{\mathrm{L}_2,\infty}^2\\&\quad+\sqrt{n}\|\mathbb{E}[g(\mathbb{E}[Y\mid X,U])\mid U]-\widehat{\mathbb{E}}[g(\widehat{\mathbb{E}}[Y\mid X,U])\mid U]\|_{\mathrm{L}_2,\infty}\|\mathbb{E}[X\mid U]-\widehat{\mathbb{E}}[X\mid U]\|_{\mathrm{L}_2}.\end{aligned}$$

Under the extra rate conditions as assumed, we further have $\|\xi\|_{2,\infty}=o_{\mathbb{P}}(1)$. This completes the proof. □

Lemma C.5.4 (Multivariate Lindeberg CLT for triangular array)**.** Let $m=m_n$ and $p=p_n$ be two sequences indexed by n. Consider the influence-function-based linear expansion for estimator $\widehat{\tau}_j$ of $\tau_j\in\mathbb{R}^d$:

$$\sqrt{n}(\widehat{\tau}_j-\tau_j)=\sqrt{n}(\mathbb{P}_n-\mathbb{P})\{\varphi_{m_nj}\}+\varsigma_{m_nj},\qquad j=1,\dots,p$$

where $\varphi_{m_n j}$ is the influence function that depends on m and the residual ς_j's satisfy that $\|\varsigma_{m_n}\|_{2,\infty} = o_{\mathbb{P}}(1)$ as $n \to \infty$. Further assume that (i) there exists a constant $c > 0$, such that $\mathbb{V}(\varphi_{m_n j}(O_1)) \geq c$, and (ii) $\max_{k\in[n]} \mathbb{E}[\|\mathbb{V}\{\varphi_{m_n j}(O_k)\}^{-\frac{1}{2}}\varphi_{m_n j}(O_k)\|^{2+\frac{2}{\delta}}] \leq M$, then

$$\sqrt{n}\mathbb{V}\{\varphi_{m_n j}\}^{-1/2}(\widehat{\tau}_j - \tau_j) \xrightarrow{\mathrm{d}} \mathcal{N}_d(0, I_d)$$

Proof of Lemma C.5.4. Note that $\varphi_{m_n j}$ is the centered influence function such that $\mathbb{E}[\varphi_{m_n j}(O)] = 0$. Let $X_{nk} = \mathbb{V}\{\varphi_{m_n j}(O_k)\}^{-\frac{1}{2}}\varphi_{m_n j}(O_k)$. From assumption (ii) that $\max_{k\in[n]} \mathbb{E}[\|X_{nk}\|^{2+\frac{2}{\delta}}] \leq M$, we have that, for any $\xi > 0$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\|X_{nk}\|^2 \mathbb{1}\{\|X_{nk}\| \geq \xi\sqrt{n}\}\right] \leq \frac{1}{n^{1+\frac{1}{\delta}}}\sum_{k=1}^{n} \mathbb{E}\left[\|X_{nk}\|^{2+\frac{2}{\delta}}\right] \leq M\frac{n}{n^{1+\frac{1}{\delta}}} \to 0.$$

This verifies Lindeberg's condition for a triangular array of random variables. From the multivariate Lindeberg's theorem (e.g., Billingsley [18, Theorem 29.5]), it follows that

$$\sqrt{n}\mathbb{V}\{\varphi_{m_n j}\}^{-\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})\{\varphi_{m_n j}\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

as $n \to \infty$. From assumption (i) that $\mathbb{V}(\varphi_{m_n j}) \geq c > 0$ and $\max_{j\in p_n} \|\varsigma_{m_n j}\| = o_{\mathbb{P}}(1)$, we further have

$$\max_{j\in p_n} \|\mathbb{V}\{\varphi_{m_n j}\}^{-\frac{1}{2}}\varsigma_{m_n j}\| = o_{\mathbb{P}}(1)$$

as $n \to \infty$. Consequently, the conclusion follows. □

Lemma C.5.5 (Matrix Chebyshev inequality)**.** Let A denote a random matrix in $\mathbb{R}^{d\times r}$ and $\beta \in \mathbb{R}^{r\times p}$ such that $\mathbb{E}[A\beta] = 0_{d\times p}$. Then with probability at least $1 - \epsilon$, it holds that

$$\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A\beta\}\|_{2,\infty} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A\|_{\mathrm{op}}^2\|\beta\|_{2,\infty}^2]^{\frac{1}{2}},$$

and

$$\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A\}\|_{\mathrm{op}} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A\|_{\mathrm{op}}^2]^{\frac{1}{2}}.$$

Proof of Lemma C.5.5. By Chebyshev inequality, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A\beta\}\|_{2,\infty} > t) &\leq \frac{n\mathbb{E}[\|(\mathbb{P}_n - \mathbb{P})\{A\beta\}\|_{2,\infty}^2]}{t^2} \\
&\leq \frac{n\mathbb{E}[\|A\beta - \mathbb{E}[A\beta]\|_{2,\infty}^2]}{nt^2} \\
&= \frac{\mathbb{E}[\|A\beta\|_{2,\infty}^2]}{t^2}.
\end{aligned}$$

Choosing $t = \mathbb{E}[\|A\beta\|_{2,\infty}^2]^{\frac{1}{2}}\epsilon^{-\frac{1}{2}}$ yields that, with probability at least $1 - \epsilon$,

$$\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A\beta\}\|_{2,\infty} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A\beta\|_{2,\infty}^2]^{\frac{1}{2}} \leq \epsilon^{-\frac{1}{2}}\mathbb{E}[\|A\|_{\mathrm{op}}^2\|\beta\|_{2,\infty}^2]^{\frac{1}{2}},$$

which finishes the proof of the first statement.

Similarly, considering all unit vectors in the unit sphere $\mathbb{S}^{r-1}$ (i.e., the set of vector $v \in \mathbb{R}^r$ such that $\|v\|_2 = 1$), it holds that

$$\mathbb{P}\left(\sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{A\}\|_{\mathrm{op}} > t\right) = \mathbb{P}\left(\sup_{v\in\mathbb{S}^{r-1}} \sqrt{n}\|(\mathbb{P}_n - \mathbb{P})\{Av\}\|_2 > t\right) \leq \frac{\mathbb{E}[\|A\|_{\mathrm{op}}^2]}{t^2}.$$

The second conclusion follows by choosing $t = (\mathbb{E}[\|A\|_{\mathrm{op}}^2]/\epsilon)^{\frac{1}{2}}$. □

C.6 Extra experimental results

C.6.1 Simulation

As in Theorem C.5.1, the nuisance functions need to be estimated fast enough such that valid inference can be guaranteed. We first examine the convergence rate of the nuisance estimations. The L_2 consistency of random forests has been examined in various studies; see, for example, [17, 153]. The rate of convergence is closely related to the minimax rate of $\mathcal{O}_{\mathbb{P}}(n^{-2/(q+2)})$ for nonparametric estimation involving q features. In a simplified setting, Biau [17] demonstrated that this rate can be improved to $\mathcal{O}_{\mathbb{P}}(n^{-0.75/(s+0.75)})$, where s represents the intrinsic dimension, which can be substantially smaller than the total feature dimension q. By numerical examination of the convergence rate for nuisance estimation, our findings indicate a L_2 convergence rate of approximately $n^{-1/4}$ for both nuisance functions on the simulated data, as illustrated in Figure C.61. This supports the appropriate use of doubly robust estimators in our experiments.

C.6.2 Real data

Extended background In a recent single-cell CRISPR perturbation study, Lalli et al. [93] investigated the molecular mechanisms of genes associated with neurodevelopmental disorders, particularly Autism Spectrum Disorder (ASD). Using a modified CRISPR-Cas9 system, they performed gene suppression experiments on 13 ASD-linked genes in the Lund Human Mesencephalic (LUHMES) neural progenitor cells. The experiment comprised 14 groups: 13 treatment groups with individual gene knockdowns and one control group. Single-cell RNA sequencing was employed to assess gene expression changes resulting from each knockdown. The authors estimated a pseudotime trajectory, which approximates the progression of neuronal differentiation. The analysis of Lalli et al. [93] suggests that some perturbations cause changes in pseudotime (slow or speed development); see Figure C.62. A scientific question of interest not answered by Lalli et al. [93] is whether some perturbation explains anything beyond the changes in expression levels caused by cell development.

In single-cell CRISPR perturbation experiments, confounding factors can significantly impact the interpretation of results. Unlike controlled experiments, these studies often resemble observational data, where confounding variables such as cell size, cell cycle stage, or microenvironment heterogeneity may influence gene expression patterns. These confounders can mask or mimic the effects of the intended genetic perturbations, potentially leading to erroneous conclusions about gene function or regulatory networks. Addressing these confounding issues is crucial for the accurate interpretation of CRISPR perturbation data and for distinguishing true biological effects from technical artifacts.

To adjust for possible confounding effects, we may take advantage of the multiple negative control genes. Even though tens of thousands of genes are measured, one typically restricts the differential expression analysis to the top thousands of highly variable genes. For the remaining genes with low variations, it is believed that there will not be sufficient power to differentiate the response from the null distribution. But even with low power, it is likely that, in total, one can detect the impact of confounding. For this reason, we use such genes as "pseudo-negative control"; even if this choice is incorrect, we still target meaningful statistical estimands, provided that the estimated embedding captures the common variability of all cells under control. Alternatively, we can also use housekeeping genes as negative control outcomes. The main goal here is to demonstrate a practical procedure for post-integrated inference and show that our asymptotic results are reasonably accurate in real data.

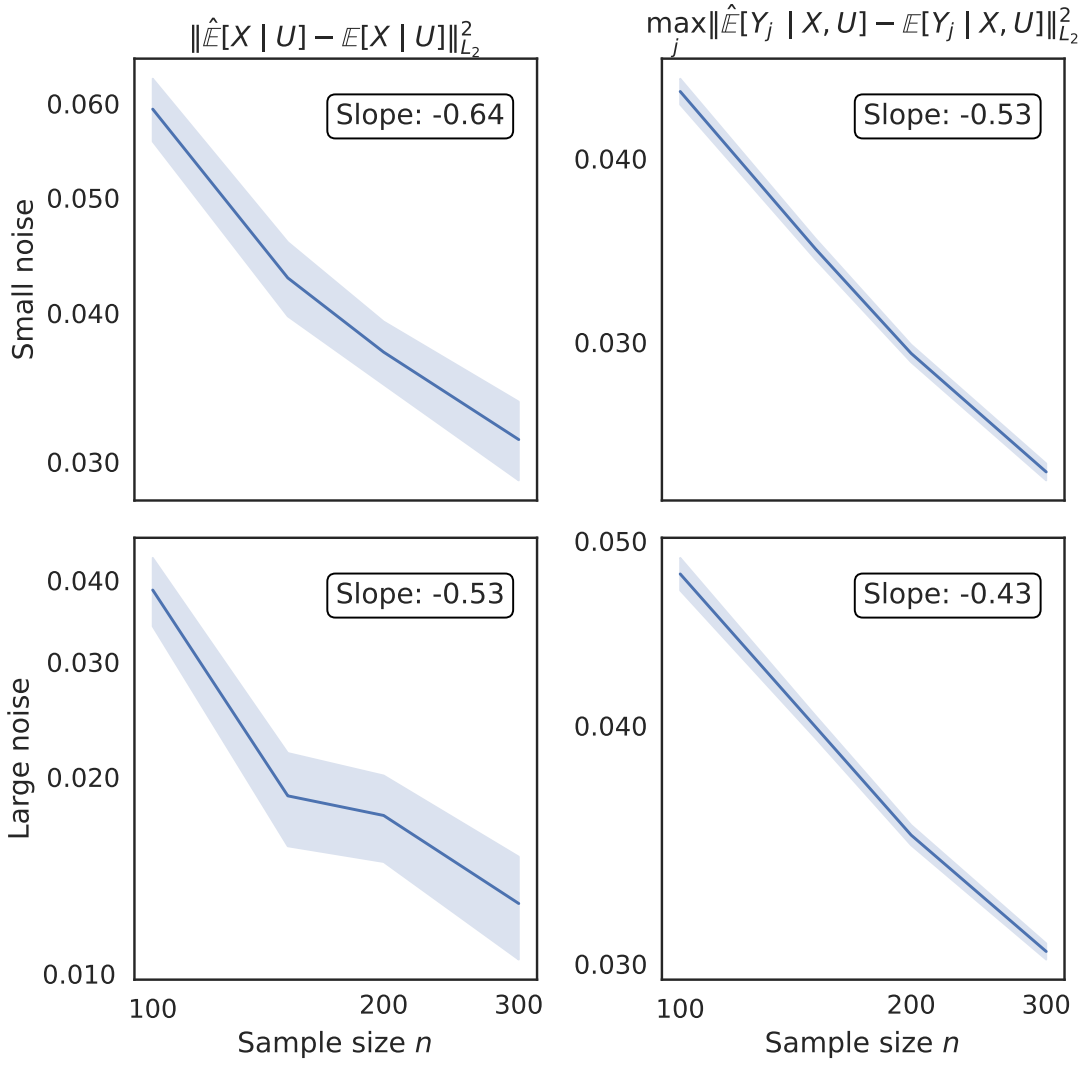

Figure C.61: Estimation error of the nuisance regression function on simulated data using random forests. The axes are shown in the logarithm scale and the slope represents the estimated rate of convergence. The data-generating process is given in Section 4.4, and we use the true latent embedding U so that the ground truth regression function is computable. The errors are computed based on 1000 test observations without irreducible additive noises.

Data. After filtering out low-quality cells and genes that expressed in less than 10 cells, we retained 8320 cells and 13086 genes under 14 perturbation conditions (including control) from Lalli et al. [93]. Following the routine selection procedure of highly variable genes in genomics [64], we select 4163 genes whose standardized variance is larger than 1, and the last 4000 genes with the lowest standardized variances are treated as negative control outcomes. The covariates we measured include the logarithm of library sizes, cell cycle scores ('S.Score' and 'G2M.Score'), batches (3 categories), and pseudotime states (normalized to range from 0 to 1). After one-hot encoding of the categorical features, we have 19 covariates (including 13 perturbation indicators), and 4163 genes for model fitting. For each highly variable gene, we aim to test whether its gene

expressions vary along the pseudotime state under perturbation conditions.

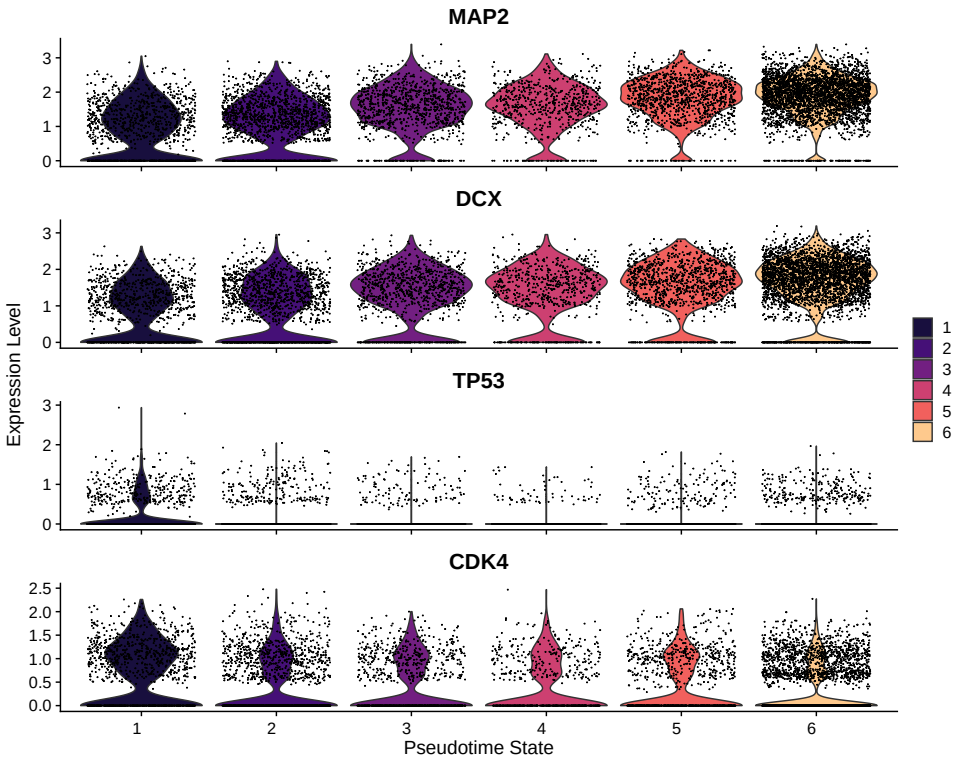

Figure C.62: Expression levels of marker genes in different estimated pseudotime states. Genes *MAP2* and *DCX* are neuronal markers (expressed in more differentiated cells) while genes *TP53* and *CDK4* are progenitor markers (expressed in less differentiated cells).

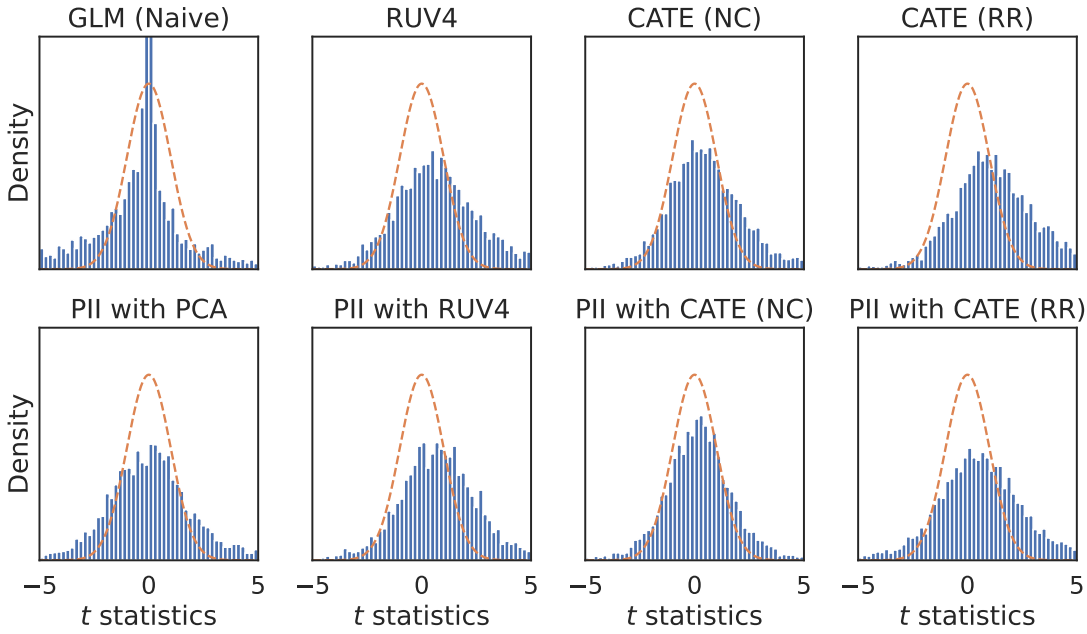

Figure C.63: Histogram of test statistics for main effects of pseudotime states on the expressions of 4163 genes. Many genes are significant because the expression levels are expected to change during neural differentiation.

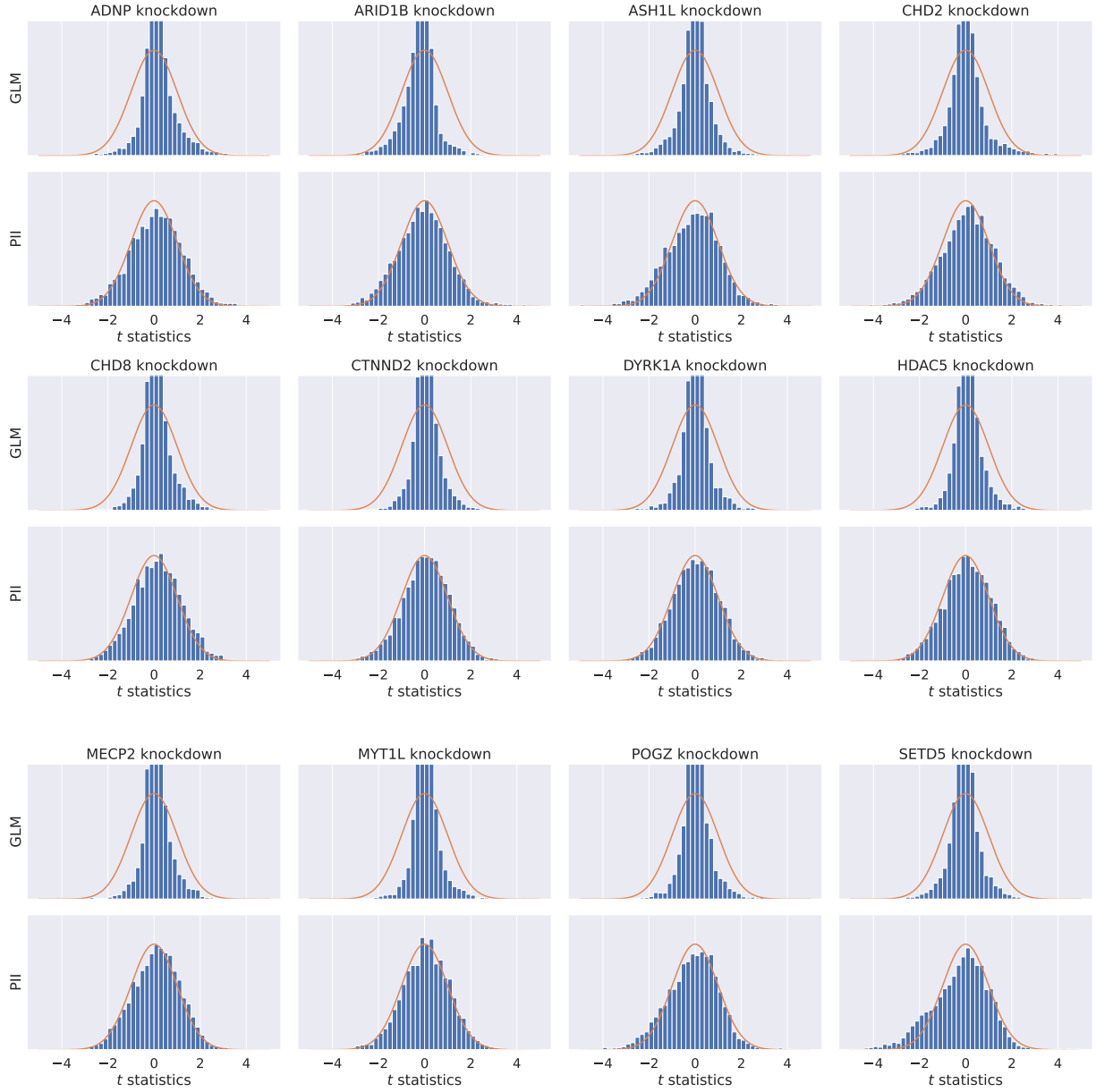

Figure C.64: Histogram of test statistics on 4163 genes for 12 different perturbation conditions. Different rows represent the results of different methods: GLM: Score tests by generalized linear models with Negative Binomial likelihood and log link function. The covariance matrix is estimated using the HC3-type robust estimator. PII: The proposed post-integrated inference with 50 principal components as the estimated embeddings.

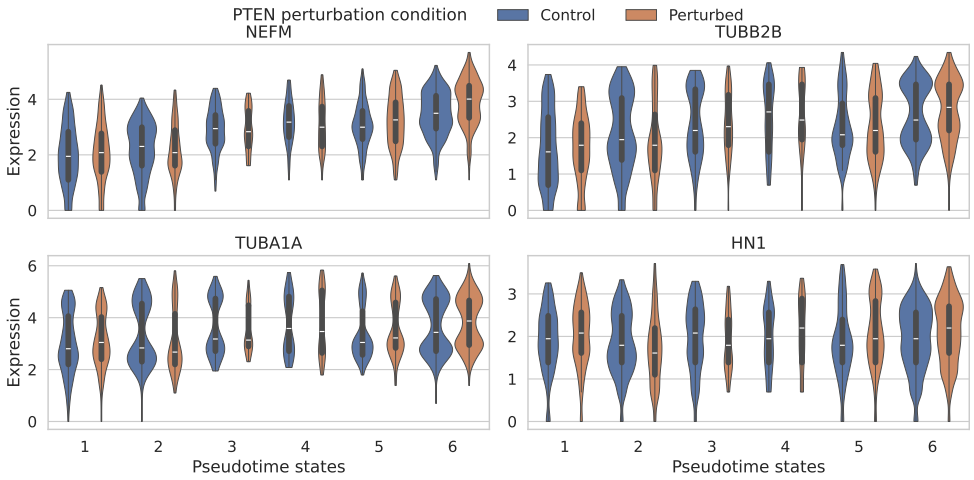

Figure C.65: Gene expressions of significant genes in the control group and the *PTEN* knockdown group. Four genes with positive estimated effect sizes are selected with a p-value threshold of 0.01 for both pseudotime states and *PTEN* knockdown for three PII methods in Figure 4.8(b) and a median expression level larger than zero.

When restricted to a small subset of significant genes discovered by PII, their expression levels are visualized as a function of pseudotime states and perturbation conditions in Figure C.65. We observe an increasing trend of the expression and the overexpression in the perturbed group at the very late stage of pseudotime. The significance suggests that these genes could be affected by not only the cell development but also the *PTEN* repression. *NEFM* is involved in neurite outgrowth and axon caliber [34], *TUBB2B* and *TUBA1A* encode critical structural subunits of microtubules that are enriched during brain development [72], *HN1* is related to cancer and senescence [75]. Given the role of *PTEN* on neural differentiation and related processes, these genes could be affected. Further research would be needed to establish any direct links between PTEN repression and the expression or function of these specific genes during neural differentiation.

Bibliography

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[2] Jeffrey Adams and Niels Richard Hansen. Substitute adjustment via recovery of latent variables. *arXiv preprint arXiv:2403.00202*, 2024.

[3] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[4] Isaiah Andrews, James H Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753, 2019.

[5] Sindri E Antonsson and Páll Melsted. Batch correction methods used in single cell rna-sequencing analyses are often poorly calibrated. *bioRxiv*, pages 2024–03, 2024.

[6] Susan Athey, Peter J Bickel, Aiyou Chen, Guido Imbens, and Michael Pollmann. Semiparametric estimation of treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2021.

[7] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

[8] Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, pages 436–465, 2012.

[9] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[10] Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023.

[11] Afonso S Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

[12] Timothy Barry, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich. Robust differential expression testing for single-cell crispr screens at low multiplicity of infection. *Genome biology*, 25(1):124, 2024.

[13] Pierre C Bellec, Jin-Hong Du, Takuya Koriyama, Pratik Patil, and Kai Tan. Corrected generalized cross-validation for finite ensembles of penalized estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae092, 2024.

[14] Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):

233–298, 2017.

[15] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv:1806.01888*, 2018.

[16] Richard Berk, Andreas Buja, Lawrence Brown, Edward George, Arun Kumar Kuchibhotla, Weijie Su, and Linda Zhao. Assumption lean regression. *The American Statistician*, 2021.

[17] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[18] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Statistics. Wiley, 1995. ISBN 9780471007104.

[19] Xin Bing, Yang Ning, and Yaosheng Xu. Adaptive estimation in multivariate response regression with hidden variables. *The Annals of Statistics*, 50(2):640–672, 2022.

[20] Xin Bing, Wei Cheng, Huijie Feng, and Yang Ning. Inference in high-dimensional multivariate response regression with hidden variables. *Journal of the American Statistical Association*, pages 1–12, 2023.

[21] T Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, pages 1–14, 2021.

[22] Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *The Journal of Machine Learning Research*, 21(1):9442–9482, 2020.

[23] Abhishek Chakrabortty, Guorong Dai, and Eric Tchetgen Tchetgen. A general framework for treatment effect estimation in semi-supervised and high dimensional settings. *arXiv preprint arXiv:2201.00468*, 2022.

[24] Jinyuan Chang, Yumou Qiu, Qiwei Yao, and Tao Zou. Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57–82, 2018.

[25] Jiahua Chen, Pengfei Li, and Yuejiao Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105, 2012.

[26] Yunxiao Chen and Xiaoou Li. Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, 109(3):769–782, 2022.

[27] Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84:124–146, 2019.

[28] Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532):1756–1770, 2020.

[29] Junyun Cheng, Gaole Lin, Tianhao Wang, Yunzhu Wang, Wenbo Guo, Jie Liao, Penghui Yang, Jie Chen, Xin Shao, Xiaoyan Lu, Ling Zhu, Yi Wang, and Xiaohui Fan. Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. *Adv Sci (Weinh)*, 10(4):e2204484, Feb 2023. doi: 10.1002/advs.202204484.

[30] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[31] Victor Chernozhukov, Christian Hansen, and Yuan Liao. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

[32] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[33] Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, 2018.

[34] Stanley KK Cheung, Jacinda Kwok, Penelope MY Or, Chi Wai Wong, Bo Feng, Kwong Wai Choy, Raymond CC Chang, J Peter H Burbach, Alfred SL Cheng, and Andrew M Chan. Neuropathological signatures revealed by transcriptomic and proteomic analysis in pten-deficient mouse models. *Scientific Reports*, 13(1):6763, 2023.

[35] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, pages 1–15, 2023.

[36] Iván Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

[37] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[38] Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature Methods*, pages 1–11, 2023.

[39] Jin-Hong Du, Zhanrui Cai, and Kathryn Roeder. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119, 2022.

[40] Jin-Hong Du, Zhanrui Cai, and Kathryn Roeder. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119, 2022. doi: 10.1073/pnas.2214414119.

[41] Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8585–8631, 23–29 Jul 2023.

[42] Jin-Hong Du, Pratik Patil, Kathryn Roeder, and Arun Kumar Kuchibhotla. Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*, 2023. doi: 10.1080/10618600.2023.2288194.

[43] Jin-Hong Du, Tianyu Chen, Ming Gao, and Jingshu Wang. Joint trajectory inference for single-cell genomics using deep learning with a mixture prior. *Proceedings of the National Academy of Sciences*, 121(37):e2316256121, 2024.

[44] Jin-Hong Du, Pratik Patil, Kathryn Roeder, and Arun Kumar Kuchibhotla. Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical*

Statistics, pages 1–12, 2024.

[45] Jin-Hong Du, Kathryn Roeder, and Larry Wasserman. Assumption-lean post-integrated inference with negative control outcomes. *arXiv preprint arXiv:2410.04996*, 2024.

[46] Jin-Hong Du, Kathryn Roeder, and Larry Wasserman. Disentangled feature importance. *arXiv preprint arXiv:2507.00260*, 2025.

[47] Jin-Hong Du, Maya Shen, Hansruedi Mathys, and Kathryn Roeder. Causal differential expression analysis under unmeasured confounders with causarray. *bioRxiv*, pages 2025–01, 2025.

[48] Jin-Hong Du, Larry Wasserman, and Kathryn Roeder. Simultaneous inference for generalized linear models with unmeasured confounders. *Journal of the American Statistical Association*, pages 1–15, 2025.

[49] Jin-Hong Du, Zhenghao Zeng, Edward H Kennedy, Larry Wasserman, and Kathryn Roeder. Causal inference for genomic data with multiple heterogeneous outcomes. *Journal of the American Statistical Association*, pages 1–14, 2025.

[50] Editorial. A focus on single-cell omics. *Nat Rev Genet*, 24(8):485, Aug 2023. doi: 10.1038/s41576-023-00628-3.

[51] Yingjie Feng. Causal inference in possibly nonlinear factor models. *arXiv preprint arXiv:2008.13651*, 2020.

[52] W Dana Flanders and Mitchel Klein. A general, multivariate definition of causal effects in epidemiology. *Epidemiology*, 26(4):481–489, 2015.

[53] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[54] Jack M Fu, F Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan L Collins, Shan Dong, Brie Wamsley, Lambertus Klei, Lily Wang, Stephanie P Hao, Christine R Stevens, Caroline Cusick, Mehrtash Babadi, Eric Banks, Brett Collins, Sheila Dodge, Stacey B Gabriel, Laura Gauthier, Samuel K Lee, Lindsay Liang, Alicia Ljungdahl, Behrang Mahjani, Laura Sloofman, Andrey N Smirnov, Mafalda Barbosa, Catalina Betancur, Alfredo Brusco, Brian H Y Chung, Edwin H Cook, Michael L Cuccaro, Enrico Domenici, Giovanni Battista Ferrero, J Jay Gargus, Gail E Herman, Irva Hertz-Picciotto, Patricia Maciel, Dara S Manoach, Maria Rita Passos-Bueno, Antonio M Persico, Alessandra Renieri, James S Sutcliffe, Flora Tassone, Elisabetta Trabetti, Gabriele Campos, Simona Cardaropoli, Diana Carli, Marcus C Y Chan, Chiara Fallerini, Elisa Giorgio, Ana Cristina Girardi, Emily Hansen-Kiss, So Lun Lee, Carla Lintas, Yunin Ludena, Rachel Nguyen, Lisa Pavinato, Margaret Pericak-Vance, Isaac N Pessah, Rebecca J Schmidt, Moyra Smith, Claudia I S Costa, Slavica Trajkova, Jaqueline Y T Wang, Mullin H C Yu, Autism Sequencing Consortium (ASC), Broad Institute Center for Common Disease Genomics (Broad-CCDG), iPSYCH-BROAD Consortium, David J Cutler, Silvia De Rubeis, Joseph D Buxbaum, Mark J Daly, Bernie Devlin, Kathryn Roeder, Stephan J Sanders, and Michael E Talkowski. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat Genet*, 54(9):1320–1331, Sep 2022. doi: 10.1038/s41588-022-01104-0.

[55] Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, Eitan S Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T Mahoney, Nick Dee, Jeff Goldy, et al. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, pages 1–18, 2024.

[56] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[57] John W Galbraith and Victoria Zinde-Walsh. Simple and reliable estimators of coefficients of interest in a model with high-dimensional confounding effects. *Journal of econometrics*, 218(2):609–632, 2020.

[58] Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

[59] David Gerard and Matthew Stephens. Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, 21(1):15–32, 2020.

[60] Karina Griesi-Oliveira and Maria Rita Passos-Bueno. Reply to lombardo, 2020: An additional route of investigation: what are the mechanisms controlling ribosomal protein genes dysregulation in autistic neuronal cells? *Mol Psychiatry*, 26(5):1436–1437, May 2021. doi: 10.1038/s41380-020-0792-7.

[61] Qiufang Guo, Yaqiong Wang, Qing Wang, Yanyan Qian, Yinmo Jiang, Xinran Dong, Huiyao Chen, Xiang Chen, Xiuyun Liu, Sha Yu, et al. In the developing cerebral cortex: axonogenesis, synapse formation, and synaptic plasticity are regulated by satb2 target genes. *Pediatric Research*, 93(6):1519–1527, 2023.

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

[63] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

[64] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

[65] Wolfgang Härdle, Enno Mammen, and Marlene Müller. Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93(444):1461–1474, 1998.

[66] Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially linear models*. Springer Science & Business Media, 2000.

[67] Kan He, Jian Zhang, Justin Liu, Yandi Cui, Leyna G Liu, Shoudong Ye, Qian Ban, Ruolan Pan, and Dahai Liu. Functional genomics study of protein inhibitor of activated stat1 in mouse hippocampal neuronal cells revealed by rna sequencing. *Aging (Albany NY)*, 13(6): 9011, 2021.

[68] Derek Hong and Lilia M Iakoucheva. Therapeutic strategies for autism: targeting three levels of the central dogma of molecular biology. *Transl Psychiatry*, 13(1):58, Feb 2023. doi: 10.1038/s41398-023-02356-y.

[69] George W Howe and C Hendricks Brown. Retrospective psychometrics and effect heterogeneity in integrated data analysis: Commentary on the special issue. *Prevention Science*, 24(8):1672–1681, 2023.

[70] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[71] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[72] Xavier Hubert Jaglin, Karine Poirier, Yoann Saillour, Emmanuelle Buhler, Guoling Tian, Nadia Bahi-Buisson, Catherine Fallet-Bianco, Françoise Phan-Dinh-Tuy, Xiang Peng Kong, Pascale Bomont, et al. Mutations in the β-tubulin gene tubb2b result in asymmetrical polymicrogyria. *Nature genetics*, 41(6):746–752, 2009.

[73] Clemens Jaitner, Chethan Reddy, Andreas Abentung, Nigel Whittle, Dietmar Rieder, Andrea Delekate, Martin Korte, Gaurav Jain, Andre Fischer, Farahnaz Sananbenesi, et al. Satb2 determines mirna expression and long-term memory in the adult central nervous system. *Elife*, 5:e17361, 2016.

[74] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[75] Qi Jia, Hongbo Nie, Peng Yu, Baiyun Xie, Chenji Wang, Fu Yang, Gang Wei, and Ting Ni. Hnrnpa1-mediated 3' utr length changes of hn1 contributes to cancer-and senescence-associated phenotypes. *Aging (Albany NY)*, 11(13):4407, 2019.

[76] Kevin Jiang and Yang Ning. Treatment effect estimation with unobserved and heterogeneous confounding variables. *arXiv preprint arXiv:2207.14439*, 2022.

[77] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):1–24, 2022.

[78] Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *arXiv preprint arXiv:2402.14264*, 2024.

[79] Xin Jin, Sean K Simmons, Amy Guo, Ashwin S Shetty, Michelle Ko, Lan Nguyen, Vahbiz Jokhi, Elise Robinson, Paul Oyler, Nathan Curry, Giulio Deangeli, Simona Lodato, Joshua Z Levin, Aviv Regev, Feng Zhang, and Paola Arlotta. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520), Nov 2020. doi: 10.1126/science.aaz6063.

[80] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[81] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25(16):1–59, 2024.

[82] Martin Kampmann. Crispr-based functional genomics for neurological disease. *Nat Rev Neurol*, 16(9):465–480, Sep 2020. doi: 10.1038/s41582-020-0373-z.

[83] Yasar Arfat T Kasu, Akshaya Arva, Jess Johnson, Christin Sajan, Jasmin Manzano, Andrew Hennes, Jacy Haynes, and Christopher S Brower. Bag6 prevents the aggregation of neurodegeneration-associated fragments of tdp43. *Iscience*, 25(5), 2022.

[84] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

[85] Edward H Kennedy, Shreya Kangovi, and Nandita Mitra. Estimating scaled treatment

effects with multiple outcomes. *Statistical methods in medical research*, 28(4):1094–1104, 2019.

[86] Edward H Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, 2020.

[87] Edward H Kennedy, Sivaraman Balakrishnan, and LA Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

[88] Kwangho Kim, Jisu Kim, and Edward H Kennedy. Causal effects based on distributional distances. *arXiv preprint arXiv:1806.02935*, 2018.

[89] Kyongwon Kim, Bing Li, Zhou Yu, and Lexin Li. On post dimension reduction statistical inference. *Annals of Statistics*, 48(3):1567–1592, 2020.

[90] Arun K Kuchibhotla and Rohit K Patra. On least squares estimation under heteroscedastic and heavy-tailed errors. *The Annals of Statistics*, 50(1):277–302, 2022.

[91] Mark J Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

[92] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21: 1–35, 2020.

[93] Matthew A Lalli, Denis Avey, Joseph D Dougherty, Jeffrey Milbrandt, and Robi D Mitra. High-throughput single-cell functional elucidation of neurodevelopmental disease–associated genes reveals convergent mechanisms altering neuronal differentiation. *Genome research*, 30(9):1317–1331, 2020.

[94] Ming-Hui Lee, Yao-Hsiang Shih, Sing-Ru Lin, Jean-Yun Chang, Yu-Hao Lin, Chun-I Sze, Yu-Min Kuo, and Nan-Shan Chang. Zfra restores memory deficits in alzheimer's disease triple-transgenic mice by blocking aggregation of trappc6aδ, sh3glb2, tau, and amyloid β, and inflammatory nf-κb activation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):189–204, 2017.

[95] Seunggeun Lee, Wei Sun, Fred A Wright, and Fei Zou. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika*, 104(2):303–316, 2017.

[96] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–35, Sep 2007. doi: 10.1371/journal.pgen.0030161.

[97] Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

[98] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–9, Oct 2010. doi: 10.1038/nrg2825.

[99] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[100] Erich Leo Lehmann and Henry Scheffé. *Completeness, similar regions, and unbiased estimation—part II*. Springer, 2012.

[101] Tenglong Li, Yuqing Zhang, Prasad Patil, and W Evan Johnson. Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. *Biostatistics*, 24(3):635–652, 2023.

[102] Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520): 1759–1769, 2017.

[103] Hui Liang, Xi Chen, Qi Yin, Danhui Ruan, Xuyang Zhao, Cong Zhang, Michael A McNutt, and Yuxin Yin. Ptenβ is an alternatively translated isoform of pten that regulates rdna transcription. *Nature communications*, 8(1):1–14, 2017.

[104] Kevin Z Lin, Jing Lei, and Kathryn Roeder. Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. *Journal of the American Statistical Association*, 116(534):457–470, 2021.

[105] Kevin Z Lin, Yixuan Qiu, and Kathryn Roeder. esvd-de: Cohort-wide differential expression in single-cell rna-seq data using exponential-family embeddings. *BMC bioinformatics*, 25 (1):113, 2024.

[106] Xihong Lin, Louise Ryan, Mary Sammel, Daowen Zhang, Chantana Padungtod, and Xiping Xu. A scaled linear mixed model for multiple outcomes. *Biometrics*, 56(2):593–601, 2000.

[107] Michael V Lombardo. Ribosomal protein genes in post-mortem cortical tissue and ipsc-derived neural progenitor cells are commonly upregulated in expression in autism. *Mol Psychiatry*, 26(5):1432–1435, May 2021. doi: 10.1038/s41380-020-0773-x.

[108] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

[109] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[110] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[111] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[112] Xiangyu Luo and Yingying Wei. Batch effects correction with unknown subtypes. *Journal of the American Statistical Association*, 2018.

[113] Monia Lupparelli and Alessandra Mattei. Joint and marginal causal effects for binary non-independent outcomes. *Journal of Multivariate Analysis*, 178:104609, 2020.

[114] Rong Ma, Eric D Sun, David Donoho, and James Zou. Principled and interpretable alignability testing and integration of single-cell data. *Proceedings of the National Academy of Sciences*, 121(10):e2313719121, 2024.

[115] Enno Mammen, Christoph Rothe, and Melanie Schienle. Nonparametric regression with

nonparametrically generated covariates. *The Annals of Statistics*, pages 1132–1170, 2012.

[116] Hansruedi Mathys, Zhuyu Peng, Carles A Boix, Matheus B Victor, Noelle Leary, Sudhagar Babu, Ghada Abdelhady, Xueqiao Jiang, Ayesha P Ng, Kimia Ghafari, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology. *Cell*, 186(20):4365–4385, 2023.

[117] Alessandra Mattei, Fan Li, and Fabrizia Mealli. Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, pages 2336–2360, 2013.

[118] José L McFaline-Figueroa, Sanjay Srivatsan, Andrew J Hill, Molly Gasperini, Dana L Jackson, Lauren Saunders, Silvia Domcke, Samuel G Regalado, Paul Lazarchuck, Sarai Alvarez, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy. *Cell Genomics*, 4(2), 2024.

[119] Chris McKennan and Dan Nicolae. Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika*, 106(4):823–840, 2019.

[120] Chris McKennan and Dan Nicolae. Estimating and accounting for unobserved covariates in high-dimensional correlated data. *Journal of the American Statistical Association*, 117 (537):225–236, 2022.

[121] Fabrizia Mealli and Barbara Pacini. Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503):1120–1131, 2013.

[122] Fabrizia Mealli, Barbara Pacini, and Elena Stanghellini. Identification of principal causal effects using additional outcomes in concentration graphs. *Journal of Educational and Behavioral Statistics*, 41(5):463–480, 2016.

[123] Andrea Mercatanti, Fan Li, and Fabrizia Mealli. Improving inference of gaussian mixtures using auxiliary variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(1):34–48, 2015.

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

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

[126] Wang Miao, Xu Shi, Yilin Li, and Eric J Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*, pages 1–12, 2024.

[127] Haeun Moon, Jin-Hong Du, Jing Lei, and Kathryn Roeder. Augmented doubly robust post-imputation inference for proteomic data. *bioRxiv*, pages 2024–03, 2024.

[128] Haeun Moon, Jin-Hong Du, Jing Lei, and Kathryn Roeder. Augmented doubly robust post-imputation inference for proteomic data. *bioRxiv*, pages 2024–03, 2025.

[129] Raffaella Nativio, Yemin Lan, Greg Donahue, Oksana Shcherbakova, Noah Barnett, Katelyn R Titus, Harshini Chandrashekar, Jennifer E Phillips-Cremins, Nancy M Bonini, and Shelley L Berger. The chromatin conformation landscape of alzheimer's disease. *bioRxiv*,

pages 2024–04, 2024.

[130] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

[131] Jing Ouyang, Kean Ming Tan, and Gongjun Xu. High-dimensional inference for generalized linear models with hidden confounding. *Journal of Machine Learning Research*, 24(296): 1–61, 2023.

[132] Art B Owen and Jingshu Wang. Bi-cross-validation for factor analysis. *Statistical Science*, pages 119–139, 2016.

[133] Susan M Paddock, Carolina Franco, F Jay Breidt, and Brenda Betancourt. Statistical data integration for health policy evidence-building. *Annual Review of Statistics and Its Application*, 12, 2024.

[134] Yongjin P Park and Manolis Kellis. Cocoa-diff: counterfactual inference for single-cell gene expression analysis. *Genome Biology*, 22(1):1–23, 2021.

[135] Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[136] Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319):1–113, 2023.

[137] Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022.

[138] Stuart J Pocock, Nancy L Geller, and Anastasios A Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, pages 487–498, 1987.

[139] Yumou Qiu, Jiarui Sun, and Xiao-Hua Zhou. Unveiling the unobservable: Causal inference on multiple derived outcomes. *Journal of the American Statistical Association*, pages 1–12, 2023.

[140] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotechnol*, 32(9):896–902, Sep 2014. doi: 10.1038/nbt.2931.

[141] James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.

[142] James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

[143] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[144] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal*

of the Econometric Society, pages 931–954, 1988.

[145] Paul R Rosenbaum. The role of known effects in observational studies. *Biometrics*, pages 557–569, 1989.

[146] Jason Roy, Xihong Lin, and Louise M Ryan. Scaled marginal models for multiple continuous outcomes. *Biostatistics*, 4(3):371–383, 2003.

[147] Agus Salim, Ramyar Molania, Jianan Wang, Alysha De Livera, Rachel Thijssen, and Terence P Speed. Ruv-iii-nb: Normalization of single cell rna-seq data. *Nucleic Acids Research*, 50(16):e96–e96, 2022.

[148] Mary Sammel, Xihong Lin, and Louise Ryan. Multivariate linear mixed models for multiple outcomes. *Statistics in medicine*, 18(17-18):2479–2492, 1999.

[149] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *R Soc Open Sci*, 9(8):220638, Aug 2022. doi: 10.1098/rsos.220638.

[150] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature genetics*, 53(6):770–777, 2021.

[151] Sergi Sayols. rrvgo: a bioconductor package for interpreting lists of gene ontology terms. *Micropublication Biology*, 2023, 2023.

[152] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[153] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, pages 1716–1741, 2015.

[154] Thomas A Severini and Joan G Staniswalis. Quasi-likelihood estimation in semiparametric models. *Journal of the American statistical Association*, 89(426):501–511, 1994.

[155] Jay Shendure, Gregory M Findlay, and Matthew W Snyder. Genomic medicine-progress, pitfalls, and promise. *Cell*, 177(1):45–57, Mar 2019. doi: 10.1016/j.cell.2019.02.003.

[156] Xu Shi, Wang Miao, Jennifer C Nelson, and Eric 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(2): 521–540, 2020.

[157] Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

[158] Jordan W Squair, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, Taha Qaiser, Kaya J E Matson, Quentin Barraud, Ariel J Levine, Gioele La Manno, Michael A Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nat Commun*, 12(1):5692, Sep 2021. doi: 10.1038/s41467-021-25960-2.

[159] Yinrui Sun, Li Ma, and Yin Xia. A decorrelating and debiasing approach to simultaneous inference for high-dimensional confounded models. *Journal of the American Statistical Association*, pages 1–12, 2023.

[160] Yunting Sun, Nancy R Zhang, and Art B Owen. Multiple hypothesis testing adjusted for

latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, pages 1664–1688, 2012.

[161] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[162] Armando Teixeira-Pinto and Laura Mauri. Statistical analysis of noncommensurate multiple outcomes. *Circulation: Cardiovascular Quality and Outcomes*, 4(6):650–656, 2011.

[163] Sally W Thurston, David Ruppert, and Philip W Davidson. Bayesian models for multiple outcomes nested in domains. *Biometrics*, 65(4):1078–1086, 2009.

[164] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[165] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

[166] Lloyd N Trefethen and David Bau. *Numerical linear algebra*. SIAM, 2022.

[167] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[168] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[169] Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[170] Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 1996. ISBN 9781475725452.

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

[172] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[173] Nico Wahl, Sergio Espeso-Gil, Paola Chietera, Amelie Nagel, Aodán Laighneach, Derek W Morris, Prashanth Rajarajan, Schahram Akbarian, Georg Dechant, and Galina Apostolova. Satb2 organizes the 3d genome architecture of cognition in cortical neurons. *Molecular Cell*, 84(4):621–639, 2024.

[174] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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

[176] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR, 2017.

[177] Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence

interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):752–776, 2021.

[178] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[179] Shu Yang, Jae Kwang Kim, and Rui Song. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):445–465, 2020.

[180] Frank B Yoon, Garrett M Fitzmaurice, Stuart R Lipsitz, Nicholas J Horton, Nan M Laird, and Sharon-Lise T Normand. Alternative methods for testing treatment effects on the basis of multiple outcomes: simulation and case study. *Statistics in medicine*, 30(16):1917–1932, 2011.

[181] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[182] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1):174, 2017.

[183] Zhenghao Zeng, Sivaraman Balakrishnan, Yanjun Han, and Edward H Kennedy. Causal inference with high-dimensional discrete covariates. *arXiv preprint arXiv:2405.00118*, 2024.

[184] Lei Zhang, Ning-Ning Song, Qiong Zhang, Wan-Ying Mei, Chun-Hui He, Pengcheng Ma, Ying Huang, Jia-Yin Chen, Bingyu Mao, Bing Lang, et al. Satb2 is required for the regionalization of retrosplenial cortex. *Cell Death & Differentiation*, 27(5):1604–1617, 2020.

[185] Mengqi Zhang, Si Liu, Zhen Miao, Fang Han, Raphael Gottardo, and Wei Sun. Ideas: individual level differential expression analysis for single-cell rna-seq data. *Genome biology*, 23(1):1–17, 2022.

[186] Zhaojun Zhang, Divij Mathew, Tristan Lim, Kaishu Mason, Clara Morral Martinez, Sijia Huang, E John Wherry, Katalin Susztak, Andy J Minn, Zongming Ma, et al. Signal recovery in single cell batch integration. *bioRxiv*, 2023.

[187] Yaoming Zhen and Jin-Hong Du. Network-based neighborhood regression. *Journal of the American Statistical Association*, pages 1–14, 2025.

[188] Wenbin Zhou and Jin-Hong Du. Distance-preserving spatial representations in genomic data. *arXiv preprint arXiv:2408.00911*, 2024.

[189] Ying Zhou, Dingke Tang, Dehan Kong, and Linbo Wang. Promises of parallel outcomes. *Biometrika*, 111(2):537–550, 2024.