Statistical Inference for Interactions and Infinite-Dimensional Estimands

Aaron Hudson

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Ali Shojaie, Chair

Kenneth Rice

Marco Carone

Program Authorized to Offer Degree:
University of Washington Department of Biostatistics

University of Washington

Abstract

Statistical Inference for Interactions and
Infinite-Dimensional Estimands

Aaron Hudson

Chair of the Supervisory Committee:
Professor Ali Shojaie
Department of Biostatistics

In this dissertation, we make methodological contributions to three statistical inference problems. We first consider two challenges that can arise when assessing for interactions, and we then discuss inference for complex estimands in semiparametric and nonparametric models. In Chapter 2, we propose an inferential procedure for identifying when an effect is substantially stronger in some sub-populations than in others. In Chapter 3, we propose a method for differential network analysis in the setting where the dependencies shared among the nodes are associated with covariates. In Chapter 4, we introduce an approach to inference on infinite-dimensional estimands that does not require the estimand of interest to take a parametric form. We conclude with a summary and a discussion of potential future research directions in Chapter 5.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

My accomplishments so far would not have been achievable without an overwhelming amount of support from family, friends, and mentors.

I am indebted to my family for their unconditional love and support. I thank my aunts, uncles, grandparents, and extended family for being an ever-reliable and enthusiastic cheer leading section. I thank my mother and father for being deeply involved in every step of my education, to the furthest extent of their ability. I also extend thanks to my older brother, Stephen. I cherish our close friendship, and I doubt that I would have the same level of ambition if he had not set such a strong example of high achievement before me.

I was fortunate to find a supportive community among students in the graduate program. The students were invested in each other's success and would often provide help and feedback in coursework and research. We had a lot of fun together, and I have fond memories of our many trips, outings, and chats. I formed wonderful friendships at the University of Washington that I sincerely hope will last.

I also owe my success to mentorship from faculty and staff members in the graduate program. In particular, I thank Gitana Garofalo, who would always offer encouragement and a listening ear to myself and all other graduate students, and would help guide us through this challenging academic maze. I thank Marco Carone and Ken Rice, who shared with me valuable insights about research, writing, and presentation, which I know will serve me well throughout my career. I also thank Noah Simon, who serves as a strong role model not only for excellence in research and teaching, but more importantly, for kindness, conscientiousness, and empathy.

I finally would like to thank my dissertation advisor, Ali Shojaie. Working with Ali was

always easy and enjoyable, and I appreciate the guidance he offered over the course of my entire graduate education. Most importantly, he never displayed any lack of confidence in my ability to succeed, even when I myself was doubtful. I would not have made it without his patience and his unwavering trust in me.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1762114. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

DEDICATION

This dissertation is dedicated to my parents, Delores and Virgil,
and to my brother, Stephen.

Chapter 1

INTRODUCTION

This dissertation consists of three projects with two major themes. The first two projects, discussed in Chapters 2 and 3, address problems that arise when performing statistical inference for interactions, that is, assessing how the associations shared among variables differ by sub-population. This work has been largely motivated by applications in differential network biology, where the objective is to identify perturbations in biological systems that are associated with disease conditions. Our work can also be applied to assess treatment effect heterogeneity in clinical trials and to assess sub-population level differences in the performance of clinical prediction models. In the final project, discussed in Chapter 4, we proprose a general framework for inference on infinite-dimensional estimands in nonparametric and semiparametric models. This work can be applied to reduce the risk of model misspecification in a wide variety of inference problems. In this chapter, we briefly describe the central problem each chapter addresses and summarize our proposed solutions.

Chapter 2 concerns statistical inference for *qualitative interactions*. Qualitative interactions occur when a treatment effect or measure of association varies in sign by sub-population. Of particular interest in many biomedical settings are absence/presence qualitative interactions, which occur when an effect is present in one sub-population but absent in another. Absence/presence interactions arise in emerging applications in precision medicine, where the objective is to identify a set of predictive biomarkers that have prognostic value for clinical outcomes in some sub-populations but not others. They also arise naturally in gene regulatory network inference, where the goal is to identify differences in networks corresponding to diseased and healthy individuals, or to different subtypes of disease; such differences lead to identification of network-based biomarkers for diseases. While the absence/presence hypoth-

esis can be of interest in many applications, we show that developing a statistical test for this hypothesis is an intractable problem. Concluding that an absence/presence interaction occurs requires evidence that in at least one sub-population, the effect is exactly null. One can only obtain such evidence under untenable assumptions To overcome this challenge, we relax the problem in a novel inference framework. We argue that as an alternative to identifying absence/presence interactions, it is reasonable to instead study the relative difference in effect size, reasoning that when the relative difference is large, an absence/presence interaction occurs. We propose an inferential framework for studying the relative difference in absolute effect size and show that our new approach only requires mild assumptions.

In Chapter 3, we present an approach to differential network analysis with covariate adjustment. Identifying differences between biological networks that correspond to disease conditions is important for understanding the underlying disease mechanisms. When the dependencies among nodes in the network depends on covariates, methods that do not adjust for covariates can detect spurious differential connections, which are induced by the effect of the covariates on both the disease condition and the genetic network. While some methods for covariate-adjusted estimation of networks have been proposed (e.g., Zhou et al., 2010; Wang and Kolar, 2014; Ha et al., 2018; Ni et al., 2019), these approaches do not emphasize hypothesis testing for assessment of statistical significance of observed differences in the network. To address this issue, we propose a covariate-adjusted hypothesis test for differential network analysis. Our proposed method assesses differential network connectivity by testing the null hypothesis that the network is the same in individuals with different disease conditions whose covariates have equal value. We introduce a general framework for estimation and inference in exponential family pairwise interaction models in the high-dimensional setting, wherein large networks are inferred using a relatively small number of observations. We show that the covariate-adjusted test exhibits improved type-1 error control compared with naïve hypothesis testing procedures that do not account for covariates. We additionally show that there are settings in which our proposed methodology provides improved power to detect differential connections.

In Chapter 4, we introduce a general approach to hypothesis testing and confidence band construction for infinite-dimensional estimands in nonparametric and semiparametric models. It is often of interest to make inference on an unknown function that is a local parameter of the data-generating mechanism, such as a density or regression function. Such estimands can typically only be estimated at a slower-than-parametric rate in nonparametric and semiparametric models, and performing calibrated inference can be challenging. In many cases, these estimands can be expressed as the minimizer of a population risk functional. We propose a general framework that leverages such representation and provides a nonparametric extension of the score test for inference on an infinite-dimensional risk minimizer. We demonstrate that our framework is applicable in a wide variety of problems. We describe how to use our approach for inference on a mean regression function under a nonparametric model and a partially additive model.

Chapter 2

STATISTICAL INFERENCE FOR QUALITATIVE INTERACTIONS

2.1 Introduction

An objective of many biomedical studies is to identify and test for *interactions*, which arise when a measure of effect or association between variables differs by sub-population. Precision medicine and genetic network inference provide examples of areas in which interactions are of interest. For instance, researchers in precision medicine seek to understand how response to treatment differs with patient characteristics. In genetics studies, it is of interest to determine how gene co-expression networks, which summarize the associations between genes, differ with phenotype.

Interactions may lack clinical or scientific significance when differences in effect are small. In addition to detecting interactions, it is important to identify which are meaningful. For example, in precision medicine, the most important differences in treatment effect may be those in which some sub-populations of patients benefit from the treatment, while other sub-populations are harmed or unaffected. Additionally, one may want to identify differences among sub-populations in the set of biomarkers that have prognostic value for a health outcome — that is, to determine whether some biomarkers are predictive of the outcome in only a subset of the full population. Genetic network inference provides another example: When comparing sub-population level gene co-expression networks, it may be of primary interest to identify pairs of genes that share an association in some sub-populations but share no association in others, or to identify pairs that have a positive association in one sub-population and a negative association another. This is known as differential network biology (Ideker and Krogan, 2012).

Such *qualitative interactions* are the focus of this chapter. Qualitative interactions occur when a measure of effect differs in sign by sub-population. We consider two types of qualitative interactions: positive/negative interactions — also known in the literature as cross-over interactions, nonremovable interactions, and disordinal interactions (de Gonzalez et al., 2007) — and absence/presence interactions — sometimes referred to as pure interactions (VanderWeele, 2019). Positive/negative interactions occur when an effect is positive in one sub-population and negative in another, and absence/presence interactions occur when the effect is present in one population but absent in another.

Our objective is to formally test for qualitative interactions, given independent samples from each sub-population. Testing for positive/negative interactions is well-studied (Gail and Simon, 1985; Piantadosi and Gail, 1993; Pan and Wolfe, 1997; Silvapulle, 2001; Li and Chan, 2006), while testing for absence/presence interactions has received substantially less attention. Naïve approaches, to be discussed in the sequel, require an untenable minimum signal strength condition — that if an effect is present in any sub-population, it is large enough to be detected with absolute certainty. No approaches exist, to the best of our knowledge, that avoid this assumption.

In this chapter, we propose a novel framework for inference about absence/presence interactions. Our proposed methodology allows for well-calibrated hypothesis testing under mild assumptions. We also introduce a numerical summary that measures the strength of absence/presence interactions, while accounting for the uncertainty associated with parameter estimation. Additionally, we describe methods for simultaneous inference about absence/presence and positive/negative interactions. The methodology we introduce provides an effective and flexible inference tool in precision medicine and genetic network analysis, as we illustrate in simulations and an analysis of breast cancer data from The Cancer Genome Atlas (TCGA).

2.2 Background

2.2.1 Notation

As we begin to formalize the problem, we first introduce some notation. We consider two sub-populations, labeled by $g \in \{1, 2\}$. Let $\theta_g \in \mathbb{R}$ denote a measure of association in sub-population g. When convenient, we write $\theta = (\theta_1, \theta_2)$. We can consider various measures of association, such as: correlation coefficients, indicating the strength of linear relationship between two variables of interest; log odds ratios, describing the relationship between predictors and a binary outcome; and log hazard ratios, describing the association between predictors and a time-to-event outcome.

We assume that given i.i.d. samples of size n_1 and n_2 from each sub-population, $n_g^{1/2}$-consistent and asymptotically normal estimates $\hat{\theta}_g$ of θ_g are available, i.e.,

$$n_g^{1/2}\left(\hat{\theta}_g - \theta_g\right) \to_d N\left(0, \sigma_g^2\right),$$

with $\sigma_g^2 > 0$ denoting the asymptotic variance. For expositional simplicity, we assume balanced sample sizes $n_1 = n_2 = n$ (key results are stated more generally in Appendix A). We also assume σ_g^2 is known, though we can instead use a consistent estimate, as is commonly done in practice.

We now formally state the null hypotheses of no positive/negative interactions and no absence/presence interaction, labeled $H_0^{\mathrm{P/N}}$ and $H_0^{\mathrm{A/P}}$, respectively:

$$H_0^{\mathrm{P/N}} : \theta_g \geq 0 \text{ for all } g \in \{1, 2\} \textbf{ or } \theta_g \leq 0 \text{ for all } g \in \{1, 2\}$$
$$H_0^{\mathrm{A/P}} : \theta_g = 0 \text{ for all } g \in \{1, 2\} \textbf{ or } \theta_g \neq 0 \text{ for all } g \in \{1, 2\}.$$

We let $\Theta_0^{\mathrm{P/N}}, \Theta_0^{\mathrm{A/P}}$ and $\Theta_1^{\mathrm{P/N}}, \Theta_1^{\mathrm{A/P}}$ denote the corresponding null and alternative regions of the parameter space, depicted in Figure 2.1. (Recall that the null region is the set of parameters such that the null hypothesis holds, and the alternative region is the complement of the null region.) The positive/negative null region is the union of the the non-negative and non-positive orthants, and the absence/presence null region is the union of all open orthants and the origin.

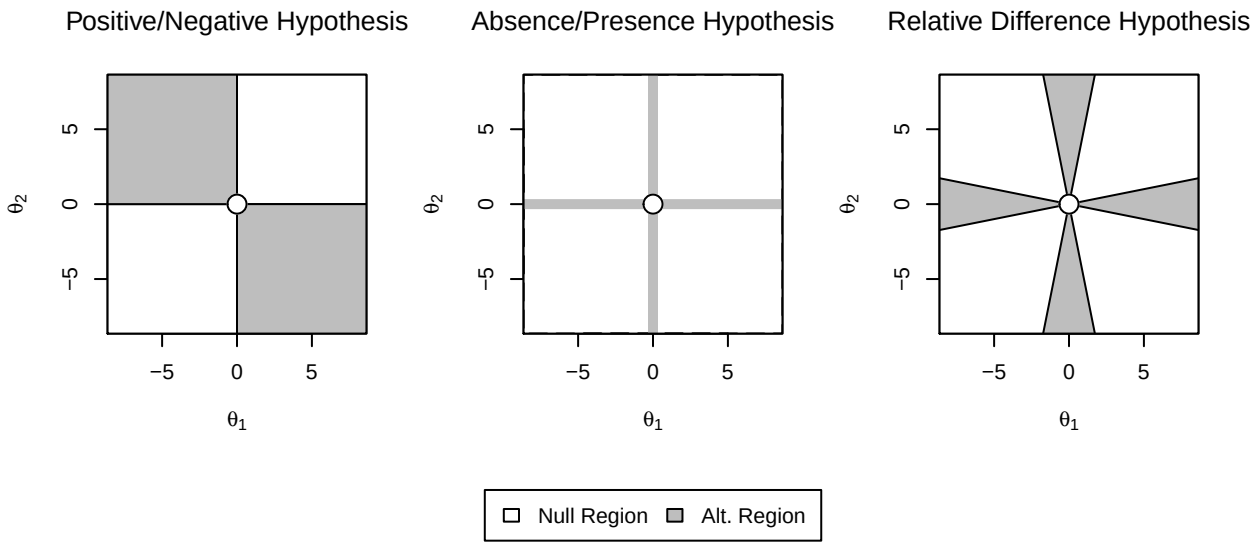


Figure 2.1: Null and alternative regions of parameter space for positive/negative, absence/presence, and relative difference hypotheses.

2.2.2 Testing composite null hypotheses

Our goal is to use the estimate $\hat{\theta}$ to perform tests of $H_0^{\mathrm{P/N}}$ and $H_0^{\mathrm{A/P}}$ such that the size is controlled asymptotically under mild assumptions. Recall that for a null hypothesis H_0 with accompanying null region Θ_0, the size of a test is defined as

$$\sup_{\theta_0 \in \Theta_0} \mathbb{P}(\text{“Reject the null hypothesis”} | \theta = \theta_0).$$

In words, the size is the largest possible type-1 error rate that could be achieved under any probability distribution given that θ belongs to the null region.

Here, we describe our general approach for controlling the size at a pre-specified level $\alpha \in (0, 1)$. We first define a test statistic T, a map from observable data to a real-valued number, with larger values of T corresponding to more evidence against the null hypothesis. We then calculate the test statistic on the observed data, which we denote by t. We write $\mathbb{P}(T > t | \theta = \theta_0)$ as the probability of observing a random test statistic at least as large as

the observed value t, assuming $\theta = \theta_0$. We reject the null hypothesis if

$$\rho(t) \equiv \sup_{\theta_0 \in \Theta_0} \mathbb{P}(T > t | \theta = \theta_0) < \alpha.$$

One can think of $\mathbb{P}(T > t|\theta = \theta_0)$ as the p-value under a specific null distribution $\theta = \theta_0$; $\rho(t)$ is then the largest of all such p-values. We reject H_0 when there is sufficient evidence to reject all hypotheses $\theta = \theta_0$. We can view $\rho(t)$ as a generalization of the usual p-value for simple null hypotheses to tests with composite null hypotheses, and will simply refer to $\rho(t)$ as "p-value". Tests of the above form are guaranteed to control the type-1 error rate at or below the pre-specified level α (Casella and Berger, 2002).

2.2.3 Existing methodology

We now review existing approaches to test for qualitative interactions. We first discuss testing positive/negative interactions before moving to absence/presence interactions.

Gail and Simon (1985) developed the most widely used procedure to test for positive/negative interactions. Though Gail and Simon proposed a general K-sample test, we focus on the two-sample problem in this chapter. We note that various K-sample tests for positive/negative interactions have been proposed (Piantadosi and Gail, 1993; Silvapulle, 2001; Li and Chan, 2006), but these procedures are essentially equivalent to the Gail-Simon test in the two-sample setting.

Gail and Simon's approach is to perform a likelihood ratio test based on the asymptotic sampling distribution of $\hat{\theta}$. The likelihood ratio test rejects $H_0^{\mathrm{P/N}}$ for large values of

$$-\frac{\sup_{(a_1,a_2)\in\Theta_0^{\mathrm{P/N}}} \prod_{g\in\{1,2\}} \sigma_g^{-1}\phi\left\{\sqrt{n}\sigma_g^{-1}(\hat{\theta}_g - a_g)\right\}}{\sup_{(b_1,b_2)\in\mathbb{R}^2} \prod_{g\in\{1,2\}} \sigma_g^{-1}\phi\left\{\sqrt{n}\sigma_g^{-1}(\hat{\theta}_g - b_g)\right\}},$$

where $\phi(\cdot)$ is the standard normal density. By performing algebraic manipulations, one can show that the likelihood ratio test equivalently rejects the null for large values of

$$T^{\mathrm{P/N}} = \min_{(a_1,a_2)\in\Theta_0^{\mathrm{P/N}}} \sum_{g\in\{1,2\}} n\left\{\sigma_g^{-1}\left(\hat{\theta}_g - a_g\right)\right\}^2.$$

The test statistic $T^{\mathrm{P/N}}$ can be interpreted as the shortest distance between $\hat{\theta}$ and the null region, where the distance is inversely weighted by the asymptotic variances of the estimates, as illustrated in Figure 2.2.

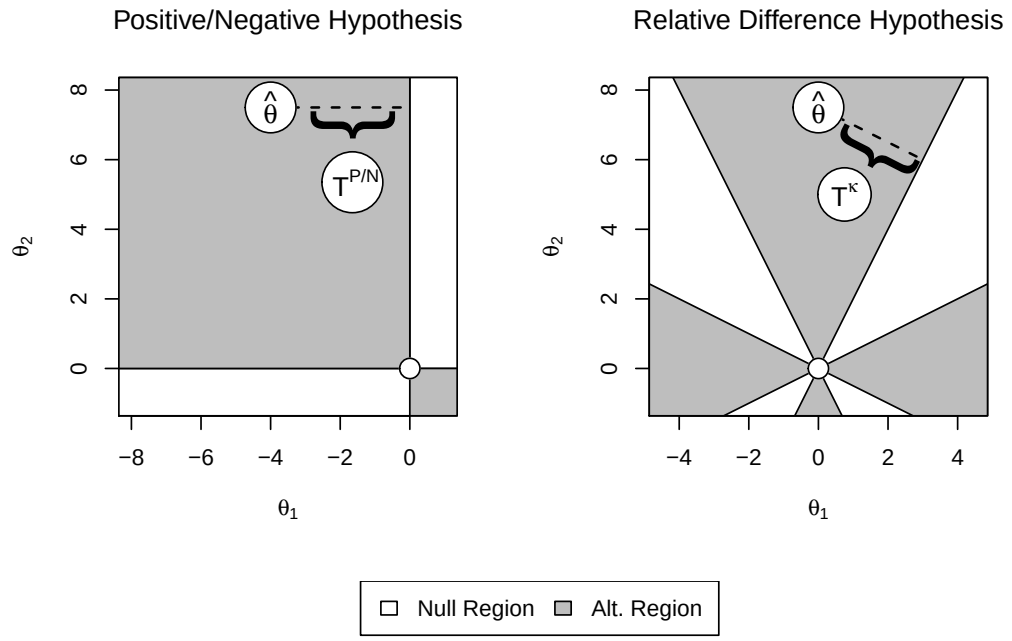


Figure 2.2: Geometric interpretation of likelihood ratio statistic for positive/negative and absence/presence hypotheses in setting where $\hat{\theta}_1$ and $\hat{\theta}_2$ have equal variance.

Gail and Simon show that the test statistic $T^{\mathrm{P/N}}$ can be calculated as

$$T^{\mathrm{P/N}} = \min_{g \in \{1,2\}} \left\{ n \left(\hat{\theta}_g / \sigma_g \right)^2 \right\} \mathbb{1} \left(\hat{\theta} \in \Theta_1^{\mathrm{P/N}} \right),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Gail and Simon also show that for $\theta \in \Theta_0$, the asymptotic tail probability $\lim_{n \to \infty} \mathbb{P}(T^{\mathrm{P/N}} > t)$ is maximized when θ_1 is infinite, and θ_2 is zero. In this case, $\hat{\theta}_1$ is strictly greater than zero with probability tending to one, and because $\hat{\theta}_2$ is asymptotically normal with mean zero, $\hat{\theta}_2$ is positive with probability tending to $1/2$. Therefore with probability tending to $1/2$, $\hat{\theta}$ is in the interior of the null region, and $T^{\mathrm{P/N}} = 0$. When $\hat{\theta}_2$ is negative, the weighted distance between $\hat{\theta}$ and the null region is $\hat{\theta}_2^2 / \sigma_g^2$.

Thus, the p-value can be calculated as

$$\rho^{\mathrm{P/N}}(t) = \sup_{\theta_0 \in \Theta_0^{\mathrm{P/N}}} \lim_{n\to\infty} \mathbb{P}\left(T^{\mathrm{P/N}} > t | \theta = \theta_0\right) = \frac{1}{2}\mathbb{P}\left(\chi_1^2 > t\right).$$

The likelihood ratio test is quite intuitive and rejects the null when a positive estimate of association is observed in one population, a negative estimate is observed in another population, and both associations are statistically significant.

Now, we discuss approaches to test for absence/presence interactions. While one might be tempted to perform a likelihood ratio test for absence/presence interactions, the likelihood ratio test fails in the sense that it *never* can reject the null. To see this, we first recognize that, similar to the positive/negative interaction test, the absence/presence likelihood ratio test would reject for large values of

$$T^{\mathrm{A/P}} = \min_{(a_1,a_2)\in\Theta_0^{\mathrm{A/P}}} \sum_{g\in\{1,2\}} n\left\{\sigma_g^{-1}\left(\hat{\theta}_g - a_g\right)\right\}^2.$$

Again, the test statistic $T^{\mathrm{A/P}}$ is the shortest distance between $\hat{\theta}$ and the null region $\Theta_0^{\mathrm{A/P}}$. Because the alternative region $\Theta_1^{\mathrm{A/P}}$ has zero area, $\hat{\theta}$ lies in the null region with probability one. Therefore, the test statistic is *always* 0, and the likelihood ratio test has no power.

One might alternatively attempt to test the absence/presence null by separately testing the null hypotheses $H_0^g : \theta_g = 0$ for $g \in \{1, 2\}$ and rejecting the absence/presence null when H_0^g is rejected for one g and not rejected for the other. To control the size of a test of this form, we need to simultaneously control the type-1 error under two scenarios: (1) there is an association in both sub-populations, and (2) there is no association in either population. When there is an association in both populations, a type-1 error occurs when we incorrectly *fail to reject* one of H_0^g. When there is no association in either population, we make a type-1 error when we incorrectly reject one of H_0^g. Thus, controlling the size of the test for $H_0^{\mathrm{A/P}}$ using this approach requires simultaneous control of the type-1 error rate and type-1I error rate of tests for H_0^g. If the tests for H_0^g are consistent — that is, the type-1I error rates tend to zero with sufficiently large samples — this approach is asymptotically valid, as only the type-1 error rates for tests of H_0^g need to be controlled. However, with even moderately

large samples, we will not be able to correctly reject false H_0^g with absolute certainty unless the true association is strong and hence easy to detect. This would make the test of $H_0^{\mathrm{A/P}}$ unreliable in the presence of weak signal.

Specifically, we require $\theta_1, \theta_2 > o(n^{-1/2})$. To see this, we construct a more formal argument. For simplicity, suppose $\sigma_1 = \sigma_2 = 1$. We consider tests of H_0^g of the form

$$\psi_g = \begin{cases} \text{Reject}; & \text{if } n^{1/2}|\hat{\theta}_g| > a \\ \text{Accept}; & \text{if } n^{1/2}|\hat{\theta}_g| < a \end{cases},$$

where a is a constant that would be selected to control the size. The probability of rejecting the absence/presence null is

$$\mathbb{P}\left(\text{Reject } H_0^{\mathrm{A/P}}\right) = \mathbb{P}\left(\psi_1 = \text{Reject}\right)\mathbb{P}\left(\psi_2 = \text{Accept}\right) + \mathbb{P}\left(\psi_1 = \text{Accept}\right)\mathbb{P}\left(\psi_2 = \text{Reject}\right).$$

Suppose θ_1 and θ_2 are guaranteed to be greater than $o(n^{-1/2})$ if they are both nonzero. Then, for $\theta_1, \theta_2 \neq 0$, $n^{1/2}|\hat{\theta}_g| \to \infty$, and $\mathbb{P}\left(\text{Reject } H_0^{\mathrm{A/P}}\right) \to 0$. Therefore, to control the size, we are only required to select α so that the type-1 error is controlled when $\theta_1 = \theta_2 = 0$; this can be done by taking a as the $(1 - \alpha/4)$ quantile of the standard normal distribution. However, if we allow $\theta_g < o(n^{1/2})$, we can see a drastically inflated type-1 error rate. For instance, for a small $\epsilon > 0$, let $\theta_1 = n^{-1/2+\epsilon}$, $\theta_2 = n^{-1/2-\epsilon}$. Then $n^{1/2}|\hat{\theta}_1| \to \infty > a$ while $n^{1/2}|\hat{\theta}_2| \to 0 < a$, so $\mathbb{P}\left(\text{Reject } H_0^{\mathrm{A/P}}\right) \to 1$. Thus, when small signal is permitted, tests of this form will be asymptotically anti-conservative.

Both approaches discussed above for testing absence/presence interactions fail for a similar reason: it is difficult to gather evidence supporting that a measure of association is exactly equal to zero. This is captured by the alternative region having zero area, causing the failure of the first approach. In the second approach, to obtain evidence supporting that an association is zero, we require that $H_{0,g}$ is only accepted when $\theta_g = 0$; for this, we rely upon a minimum signal strength condition to guarantee that any non-zero association is detected.

2.3 Proposed methodology

2.3.1 Refinement of absence/presence hypothesis

To mitigate the challenges described in Section 2.2, we consider a refinement of the absence/presence null hypothesis. The key idea is that in practice, absence/presence interactions can be approximated by considering the settings where an association is at least moderately large in one population and negligible or *near* zero in the other; or when one association is substantially stronger than the other. This means that we can expand the alternative region to include neighborhoods of zero in a way that the absence/presence interpretation is preserved.

Recall that when there exists an absence/presence interaction, the ratio of the maximum of the absolute value of the θ_g to the minimum is infinite. We cannot test that the ratio is infinite because we will never have evidence to support that the denominator is exactly zero. However, we can test that the ratio is *large* because we may have evidence to support that the denominator is very small. Motivated by this intuition, we propose to test whether the relative difference between sub-population measures of association is greater than a large pre-specified constant $\kappa > 1$. Formally, let $\theta_{\max} = \max_g |\theta_g|$ and $\theta_{\min} = \min_g |\theta_g|$. We define the new relative difference null hypothesis H_0^κ as

$$H_0^\kappa : \theta_{\max}/\theta_{\min} \leq \kappa \text{ or } \theta_{\max} = \theta_{\min} = 0.$$

Equivalently,

$$H_0^\kappa : \theta_{\max} - \kappa\theta_{\min} \leq 0. \tag{2.1}$$

The null region Θ_0^κ, illustrated in Figure 2.1, is the union of four linear subspaces — each residing in a separate orthant of $\mathbb{R}^2$; the boundary of each subspace is the union of the spans of vectors with absolute direction $(\kappa, 1)^\top$ and $(1, \kappa)^\top$.

The relative difference null region can be viewed as a relaxation of the absence/presence null. For a large choice of κ, both our original and refined null hypotheses have the same

interpretation: the greater measure of association is substantially larger than the lesser. However, we find it appealing that H_0^κ has a reasonable interpretation for any choice of κ; that is, the multiplicative difference in strength of association is no larger than κ.

To motivate defining the refined null hypothesis in terms of relative differences rather than absolute differences, we argue that testing for relative differences has at least the following benefits. First, relative differences are unitless, so the relative difference null is compatible with unitless measures of association such as the Pearson correlation coefficient, which are often preferred in the analysis of biological data. Second, a reasonable κ can be selected without prior knowledge of ranges of strength of association.

Of course, the relative difference null hypothesis depends on the choice of κ. For small values of κ, the relative difference null may be too dissimilar from the absence/presence null to retain its interpretation; for large κ, the alternative region becomes very small, and it may be difficult to detect absence/presence interactions (i.e., we may require a large sample size or a strong signal in the sub-population where an association is present). In the following subsections, we first construct a test of the relative difference null hypothesis for a pre-specified κ and then describe an approach to identify the set of κ such that the test rejects the null hypothesis, so as to circumvent tuning parameter selection.

2.3.2 Likelihood ratio test for relative difference hypothesis

We now develop a testing procedure for the new relative difference null hypothesis. The relative difference null region, unlike the absence/presence null region, has non-zero area, so a likelihood ratio test will not fail in the same manner as the likelihood ratio test for absence/presence interactions. Similar to the previously discussed examples, the likelihood ratio test statistic is

$$T^\kappa = \min_{(a_1,a_2)\in\Theta_0^\kappa} \sum_{g\in\{1,2\}} n \left\{ \sigma_g^{-1} \left(\hat{\theta}_g - a_g \right) \right\}^2,$$

and can be interpreted as the shortest (weighted) distance between $\hat{\theta}$ and the null region Θ_0^κ. Clearly, the test statistic is zero whenever $\hat{\theta}$ lies in the null region. Otherwise, T^κ is the

shortest distance between $\hat{\theta}$ and the closest of the four linear subspaces that define Θ_0^κ. The test statistic can be calculated as the distance between $(\hat{\theta}_{\max}, \hat{\theta}_{\min})^\top$ and its projection onto the span of the vector $(\kappa, 1)^\top$. The test statistic's geometric interpretation is illustrated in Figure 2.2.

The likelihood ratio test statistic is straightforward to calculate. Let $\hat{\theta}_{\max} = \max_g |\hat{\theta}_g|$ and $\hat{\theta}_{\min} = \min_g |\hat{\theta}_g|$ be the strongest and weakest estimated absolute association. In the following lemma, we state that T^κ is equal to the difference between $\hat{\theta}_{\max}$ and $\kappa\hat{\theta}_{\min}$ divided by a normalizing constant. Thus, the test statistic can be viewed as a plug-in of $\hat{\theta}$ into (2.1) with an additional normalizing constant and closely resembles the test statistic proposed by Fieller (1940) to conduct inference about ratios of means.

Lemma 2.1. *The likelihood ratio test statistic T^κ can be written as*

$$T^\kappa = \frac{\hat{\theta}_{\max} - \kappa\hat{\theta}_{min}}{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}},$$

where $\hat{\tau}_{\max} = n^{-1}\sigma_1^2\mathbb{1}\left(|\hat{\theta}_1| = \hat{\theta}_{\max}\right) + n^{-1}\sigma_2^2\mathbb{1}\left(|\hat{\theta}_2| = \hat{\theta}_{\max}\right)$, *and* $\hat{\tau}_{\min} = n^{-1}\sigma_1^2\mathbb{1}\left(|\hat{\theta}_1| = \hat{\theta}_{\min}\right) + n^{-1}\sigma_2^2\mathbb{1}\left(|\hat{\theta}_2| = \hat{\theta}_{\min}\right)$.

We now discuss how to obtain a p-value for the relative difference hypothesis. First, we obtain an observed test statistic t, a realization of T^κ calculated from the data. Following the approach described in Section 2.2, we define the p-value as

$$\rho^\kappa(t) = \sup_{\theta_0 \in \Theta_0^\kappa} \lim_{n\to\infty} \mathbb{P}(T^\kappa > t | \theta = \theta_0),$$

the largest of all asymptotic tail probabilities such that θ belongs to the null region. To determine the maximum tail probability, we characterize the limiting distribution of T^κ assuming $\theta = \theta_0$ for all θ_0 in the null region.

Though the null region contains an infinite number of values, T^κ can only attain one of three limiting distributions corresponding to the following three cases:

1. The true association is in the interior of the null region, i.e., $\theta_{\max} - \kappa\theta_{\min} < 0$.

2. The true association is on the boundary of the null region, but both associations are non-zero, i.e., $\theta_{\max} - \kappa\theta_{\min} = 0$, $\theta_{\max} > 0$, and $\theta_{\min} > 0$.

3. The true association is zero in both sub-populations, i.e., $\theta_1 = \theta_2 = 0$.

In Proposition 2.1, we describe the asymptotic behavior of T^κ for cases 1 and 2 above. We provide here some intuition for the result and reserve a formal argument for Appendix A. In case 1, because $\hat{\theta}$ is consistent, T^κ tends to minus infinity and therefore never provides evidence against the null. In case 2, T^κ asymptotically follows a standard normal distribution. To see this, we note that because both associations are non-zero, consistency and asymptotic normality of $\hat{\theta}$ imply that, for large n, the sign and order of the estimates are deterministic. That the signs are asymptotically deterministic implies that $|\hat{\theta}|$ is asymptotically normal (speaking loosely, $|\hat{\theta}_g| \to \text{sign}(\theta_g)\hat{\theta}_g$), and that order is asymptotically deterministic implies that $\hat{\theta}_{\max}$ and $\hat{\theta}_{\min}$ are asymptotically independent. Therefore, taking the difference between $\hat{\theta}_{\max}$ and $\hat{\theta}_{\min}$, suitably standardized, is asymptotically equivalent to taking a difference between two independent normal random variables with equal means. Dividing by the asymptotic variances gives the claimed result.

Proposition 2.1. *In the interior of the null region, i.e., when $\theta_{\max} - \kappa\theta_{\min} < 0$, T^κ converges in distribution to $-\infty$. At all nonzero boundary points of the null region, i.e., when $\theta_{\max} = \kappa\theta_{\min} > 0$, T^κ converges in distribution to a standard normal random variable.*

In case 3, when both associations are zero, the asymptotic distribution of the test statistic is more complicated. More specifically, $\theta_g = 0$ implies that, asymptotically, $|\hat{\theta}_g|$ follows a half-normal distribution instead of a normal distribution. Moreover, the order of $|\hat{\theta}|$ remains random in the limit. This gives rise to a non-standard limiting distribution. In particular, unlike cases 1 and 2, the limiting distribution depends on the asymptotic variances of the sub-population estimates (and also the ratio of the sample sizes in the unbalanced case). We are nonetheless able to derive an analytic expression for the distribution function, stated in Proposition A.2 (we reserve this statement for Appendix A, as the expression is cumbersome).

By Propositions 2.1 and A.2, we can calculate the p-value as the maximum of the tail probabilities in cases 2 and 3. That is,

$$\rho^{\kappa}(t) = \max\left\{1 - \Phi(t), 1 - F^{\kappa}(t)\right\}, \tag{2.2}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function and $F^{\kappa}(\cdot) \equiv \lim_{n\to\infty} \mathbb{P}(T^{\kappa} < t|\theta = \theta_0)$ is the limiting distribution function for the test statistic when $\theta = 0$. Though the limiting distribution under $\theta = 0$ is non-standard, tail probabilities can be calculated easily, as we show in Appendix A (Remark A.1). The p-value is therefore simple to calculate.

We have derived an analytic approximation for the power of the likelihood ratio test. In Proposition A.2, we more generally characterize the limiting distribution of the test statistic under hypotheses of the form $(\theta_1, \theta_2) = n^{-1/2}(c_1, c_2)$. The local asymptotic power of the proposed test at the level α is available by considering $\beta_{\alpha}^{\kappa}(c_1, c_2) \equiv \lim_{n\to\infty} \mathbb{P}\left(T^{\kappa} > t_{1-\alpha}^{*}|\theta = n^{-1/2}(c_1, c_2)\right)$, where we define $t_{1-\alpha}^{*}$ as the maximum of the $1-\alpha$ quantiles of the limiting distribution of T^{κ} under scenarios 2 and 3 described above. An analytic finite-sample approximation of the power can be calculated as $\beta_{\alpha}^{\kappa}\left(n^{1/2}\theta_1, n^{1/2}\theta_2\right)$.

A contour plot of the local asymptotic power is given in Figure 2.3. We consider both cases of equal and unequal variance of the estimators. We observe that the likelihood ratio test has low power when the strongest effect $\max\{|c_1|, |c_2|\}$ is small, and power improves considerably when the strongest effect grows. Additionally, we find that in the presence of unequal variance, the test has greater power when the weakest effect is estimated with higher precision than the strongest effect.

2.3.3 Quantifying relative difference in effect by inverting likelihood ratio test

Rather than perform a the likelihood ratio test for a pre-specified κ, one may prefer to directly estimate the relative difference in effect. Naïvely, one might consider estimating the relative difference in effect as $\hat{\theta}_{\max}/\hat{\theta}_{min}$. However, this estimate will behave poorly when $\theta_1 = \theta_2 = 0$ and should therefore not be reported in practice.

To overcome this issue, we propose to quantify the relative difference in effect size by

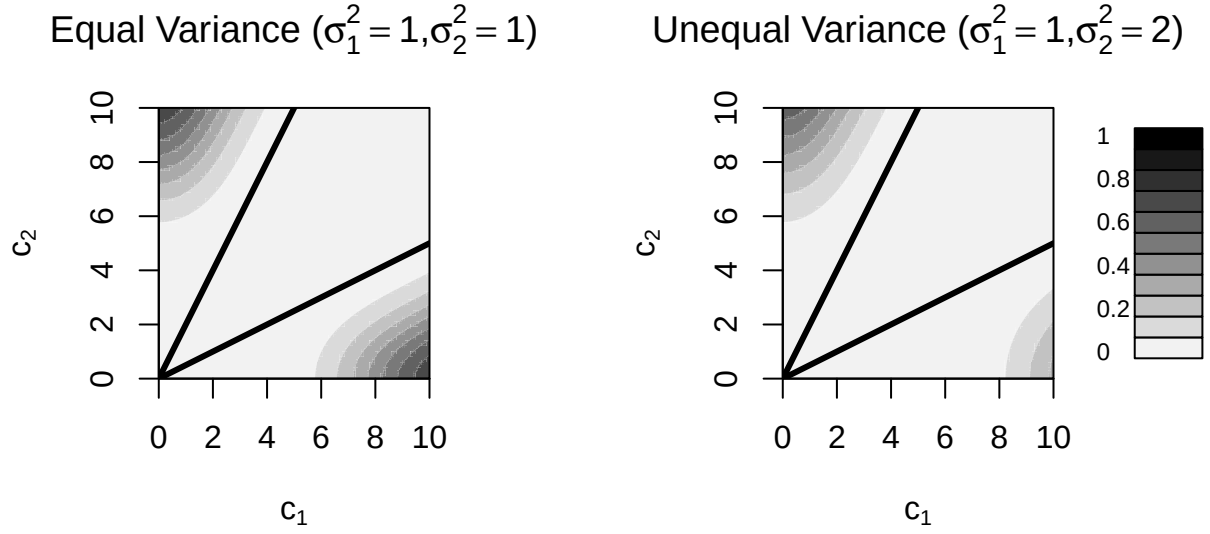


Figure 2.3: Contour plot of local asymptotic power of relative difference likelihood ratio test with $\kappa = 2$ and $\alpha = .05$. Bold lines represent the boundary of the null region. Settings of equal and unequal asymptotic variance of estimators are represented.

inverting the likelihood ratio test, similar to Fieller (1940). We define

$$\kappa^{\alpha}_{\max} \equiv \sup\{\kappa : \kappa > 1, \rho^{\kappa}(t) < \alpha\}$$

as the largest $\kappa > 1$ such that the likelihood ratio test rejects the null hypothesis at the α level. When the likelihood ratio test fails to reject for all $\kappa > 1$, we will use the convention $\kappa^{\alpha}_{\max} = 1$. We find it appealing that $\kappa^{\alpha}_{\max}$ converges to 1 if $\theta_{\max} = \theta_{\min}$, and $\kappa_{\max}$ should approach but not exceed $\theta_{\max}/\theta_{\min}$ otherwise. We discuss calculation of $\kappa^{\alpha}_{\max}$ in Appendix A.

2.3.4 Simultaneous test of qualitative interactions

When it is of interest to identify both absence/presence and positive/negative qualitative interactions, it may be desirable to test for both simultaneously. In this section, we construct an omnibus test that controls the size asymptotically.

We define the omnibus qualitative interaction null hypothesis as

$$H_0^{\mathrm{P/N},\kappa} : \text{ Both } H_0^{\mathrm{P/N}} \text{ and } H_0^{\kappa} \text{ hold.}$$

The null region $\Theta_0^{\mathrm{P/N},\kappa}$ is the intersection of the positive/negative and relative difference null regions, as depicted in Figure 2.4. We observe that as $\kappa \to \infty$, the omnibus null and alternative regions tend to the positive/negative null and alternative regions.

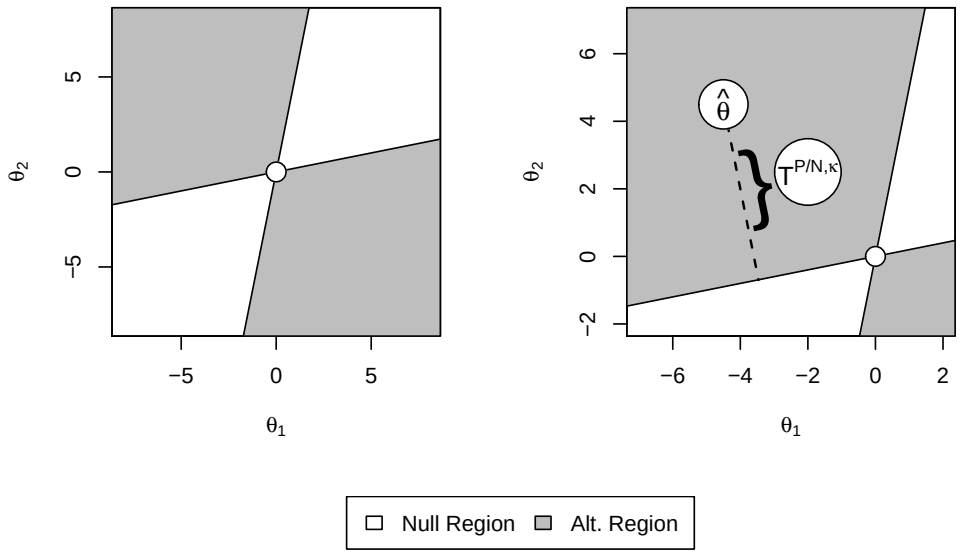


Figure 2.4: (Left) Null and alternative region for omnibus qualitative interaction hypothesis. (Right) Geometric interpretation of likelihood ratio statistic for omnibus qualitative interaction hypothesis.

To construct the likelihood ratio test, we proceed using similar arguments to those presented in Section 2.3.2. The likelihood ratio statistic $T^{\mathrm{P/N},\kappa}$ is the distance between the estimate $\hat{\theta}$ and its projection onto the null region $\Theta_0^{\mathrm{P/N},\kappa}$, inversely weighted by the asymptotic variance of $\hat{\theta}$. A simple expression for $T^{\mathrm{P/N},\kappa}$ is given in Lemma 2.2.

Lemma 2.2. *The likelihood ratio statistic $T^{\mathrm{P/N},\kappa}$ can be written as*

$$T^{\mathrm{P/N},\kappa} = \min\left\{ \frac{\left(\hat{\theta}_1 - \kappa\hat{\theta}_2\right)^2}{n^{-1}(\sigma_1^2 + \kappa^2\sigma_2^2)}, \frac{\left(\kappa\hat{\theta}_1 - \hat{\theta}_2\right)^2}{n^{-1}(\kappa^2\sigma_1^2 + \sigma_2^2)} \right\} \mathbb{1}\left(\hat{\theta} \in \Theta_1^{\mathrm{P/N},\kappa}\right).$$

Unsurprisingly, the likelihood ratio statistic for the omnibus test approaches the likelihood ratio statistic for the Gail-Simon likelihood ratio statistic for positive/negative interactions in the limit of large κ. The tests will, therefore, be nearly identical for sufficiently large κ.

To characterize the asymptotic behavior of the omnibus test statistic at each location under the null, we use similar arguments to those in Section 2.3.2. If θ belongs to the interior of the null region, $T^{\mathrm{P/N},\kappa}$ converges in probability to zero. If θ belongs to the boundary of the null region and is non-zero, $T^{\mathrm{P/N},\kappa}$ converges weakly to a uniform mixture of zero and the chi-squared distribution with one degree of freedom. If θ is zero, the limiting distribution of $T^{\mathrm{P/N},\kappa}$ is non-standard, though it can be characterized nonetheless. Formal statements of asymptotic properties of $T^{\mathrm{P/N},\kappa}$ are given in Propositions 2.2 and Remark A.2 (reserved for Appendix A).

Proposition 2.2. *If θ belongs to the interior of the null region $\Theta_0^{P/N,\kappa}$, $T^{\mathrm{P/N},\kappa}$ converges in distribution to zero. If θ is on the boundary of the null region, but $\theta \neq 0$, $\mathbb{P}(T^{\mathrm{P/N},\kappa} > t) \to \frac{1}{2}P\left(\chi_1^2 > t\right)$ as $n \to \infty$.*

Calculating the p-value for the omnibus test is no more difficult than calculating the p-value for the absence/presence test. Defining $F^{\mathrm{P/N},\kappa}(t) \equiv \lim_{n\to\infty} \mathbb{P}(T^{\mathrm{P/N}} < t|\theta = 0)$, the p-value can be calculated as

$$\rho^{\kappa,\mathrm{P/N}}(t) = \max\left\{\mathbb{P}\left(\chi_1^2 > t\right), 1 - F^{\mathrm{P/N},\kappa}(t)\right\}, \tag{2.3}$$

where t is the value of the test statistic $T^{\mathrm{P/N},\kappa}$ calculated on the observed data. We characterize the local asymptotic power of the omnibus likelihood ratio test in Proposition A.4 in Appendix A.

2.4 Simulation study

In a Monte Carlo simulation study, we examine how type-1 error rates and power of the likelihood ratio tests for qualitative interactions are affected by signal strength, sample size, and selection of κ. Additionally, we examine how $\kappa^{\alpha}_{\max}$ depends on the true sub-population effects and the sample size.

We generate random observations (Y_g, X_g) in sub-population g under the linear model:

$$Y_g = \theta_g X_g + \epsilon;\ X_g \sim N(0,1);\ \epsilon \sim N(0,1).$$

Here, X_g is the predictor of interest, Y_g is the response, and ϵ is white noise. The measure of association in which we are interested is the regression coefficient θ_g. We fix $\theta_1 = 1$ and consider $\theta_2 \in \{-1, -.9, ..., .9, 1\}$. A total of 5000 synthetic data sets are randomly generated for each θ_2 and $n \in \{50, 100\}$.

For each synthetic data set, we perform both the relative difference likelihood ratio test and the omnibus qualitative interaction likelihood ratio test with $\kappa = 2$ and $\kappa = 4$ using a significance level of $\alpha = .05$. We compare with a test for quantitative interactions based on the test statistic

$$T = \frac{(\hat{\theta}_1 - \hat{\theta}_2)^2}{n^{-1}(\sigma_1^2 + \sigma_2^2)}.$$

We additionally calculate $\kappa^{\alpha}_{\max}$ with $\alpha = .05$. Parameter estimation is performed with ordinary least squares, and model-based estimates of the standard error are used.

In Figure 2.5, we plot the Monte Carlo estimate of the rejection probability of the relative difference likelihood ratio test over the range of θ_2. We see that the test achieves control of size for both choices of κ and both sample sizes. We observe that power is largest when $|\theta_2|$ is near zero, as we would expect. Power is moderately strong when $\kappa = 2$ but fairly poor when $\kappa = 4$. With $\kappa = 4$, the likelihood ratio test has almost zero power when the sample size is small, though we see modest improvement when the sample size is larger. The test for quantitative interactions is well-powered when θ_2 is nearly zero, but the rejection probability greatly exceeds α when θ_2 is even moderately large and reasonably close to

θ_1. This illustrates that while tests for quantitative interactions may be more powerful than tests for qualitative interactions, tests for quantitative interactions are unsuitable when small between-group differences in strength of association are not of scientific interest.

Figure 2.5 also shows the estimated rejection probabilities of the omnibus test over θ_2. Control of size is achieved in both large and small samples, as expected. We note that for the omnibus test, power increases as θ_2 tends to -1, and power is uniformly larger when $\kappa = 2$ than when $\kappa = 4$.

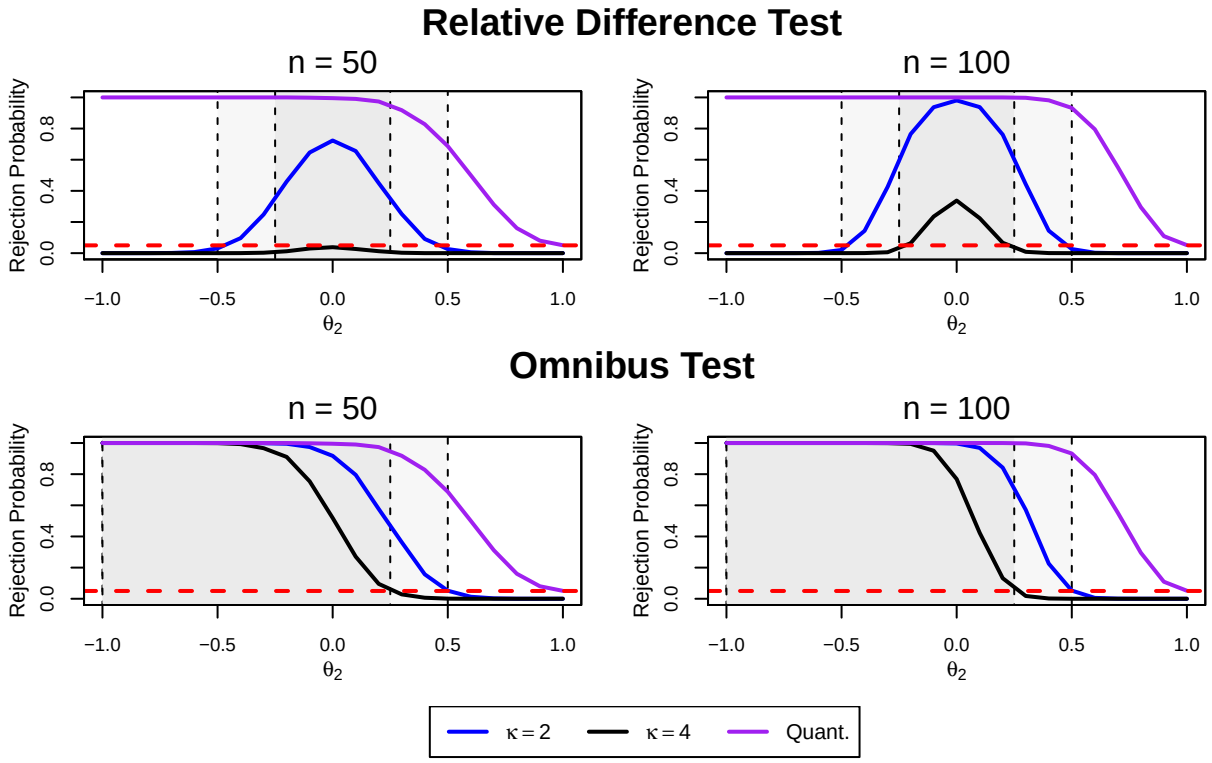


Figure 2.5: Monte Carlo estimate of rejection probability of the relative difference and omnibus likelihood ratio tests. The shaded light grey and dark grey areas correspond to sets of θ_2 such that the null hypothesis holds with $\kappa = 2$ and $\kappa = 4$, respectively. The dashed red line denotes the specified size $\alpha = .05$.

In Figure 2.6, we plot the quantiles of the $\kappa^{\alpha}_{\max}$ values from 5000 synthetic data sets for

each θ_2. We expect that for a fixed θ_2, most $\kappa^{\alpha}_{\max}$ should approach but not exceed $\theta_1/|\theta_2|$ as sample size increases; our simulations are consistent with this expectation. We note that when the sample size is small, $\kappa^{\alpha}_{\max}$ tends to underestimate the relative difference in effect size.

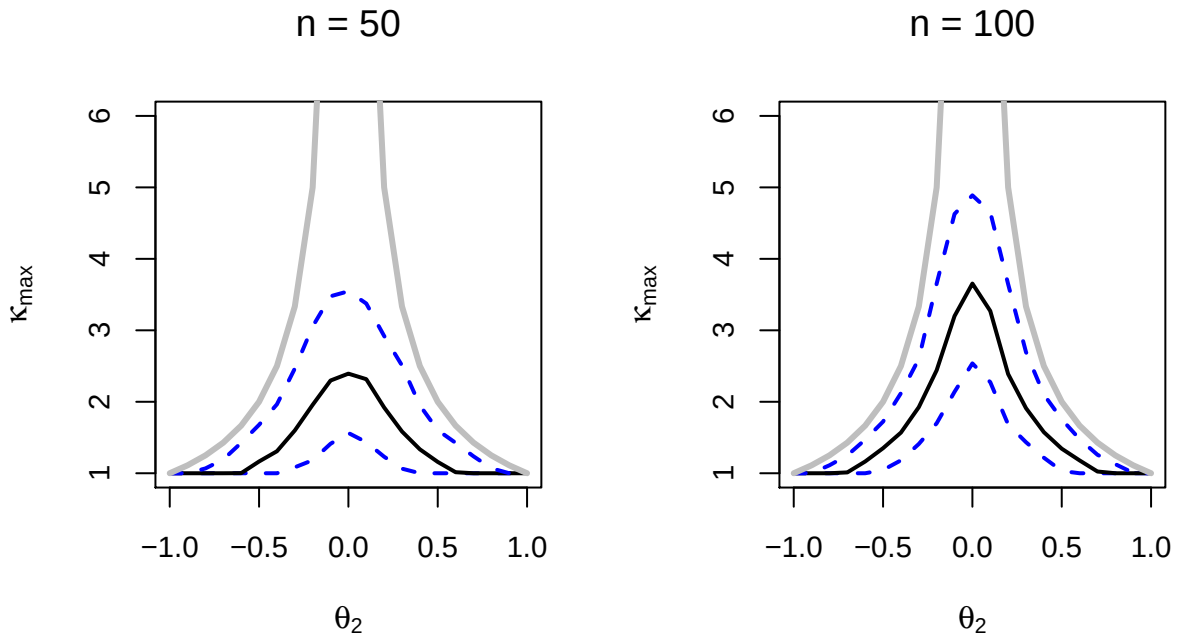


Figure 2.6: Distribution of $\kappa^{\alpha}_{\max}$ calculated on synthetic data. The median is represented by black line, and the .1 and .9 quantiles are represented by the dotted blue lines. The grey curve represents $|\theta_1|/|\theta_2|$, the value $\kappa^{\alpha}_{\max}$ is expected to approach for a given θ_2.

2.5 Data example

In this example, we investigate genetic differences in breast cancer sub-types. Classification of breast cancer based on expression of estrogen receptor (ER) is known to be associated with clinical outcomes. Approximately 70% of breast cancers are estrogen receptor positive (ER+) cancers, meaning that estrogen causes cancer cells to grow (Lumachi et al., 2013); breast cancers are otherwise estrogen receptor negative (ER-). Patients with ER+ breast

cancer tend to experience better clinical outcomes than ER- patients (Carey et al., 2006).

We conduct an analysis using publicly available data from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). We use clinical data and gene expression data from a total of 806 ER+ patients and 237 ER- patients.

We first investigate the differences between the genetic networks in ER+ and ER- breast cancer. Both ER+ and ER- breast cancer are expected to have similar pathways, but identifying differences between them may be key to understanding the underlying disease mechanism. We then conduct an analysis to assess whether any genes in a set known to be associated with breast cancer are strongly prognostic of disease outcomes in only one of the estrogen receptor groups.

2.5.1 Differential network analysis

Our objective is to determine whether there are any pairs of genes that are much more strongly associated in one estrogen receptor group than the other. We consider the set of $p = 145$ genes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) breast cancer pathway and measure the association between gene expression levels using the Pearson correlation.

We test the relative difference null hypothesis for each pair of genes with $\kappa = 2$. In Figure 2.7, we display the pairs of genes that are statistically significant at the $\alpha = .10$ level after a Bonferroni adjustment. We find that each of the genes progesterone (PGR), insulin-like growth factor 1 (IGF1R), and estrogen receptor 1 (ESR1) have multiple differential connections; each belongs to at least two pairs such that the association is twice as strong in the ER+ population than in the ER- population. These genes have been shown in the literature to be associated with sub-type and prognosis (Farabaugh et al., 2015; Reinert et al., 2019; Kurozumi et al., 2017).

2.5.2 Prognostic value of biomarkers

The goal of this analysis is to assess whether any of the KEGG genes have a stronger association with time to death in one estrogen receptor group than in the other. For each gene, we fit a univariate Cox proportional-hazards model with time to death as the outcome in both of the estrogen receptor groups separately; we measure association using the log hazard ratio. A total of 64 deaths occurred in the ER+ group, and 33 deaths occurred the ER- group. We calculate $\kappa^{\alpha}_{\max}$ with $\alpha = .10$ for each gene.

In Figure 2.7, we compare the log hazard ratios of the ER+ and ER- groups in a scatterplot. Though the log hazard ratios for most genes are similar between subgroups, there are twelve genes with $\kappa^{\alpha}_{\max}$ larger than one. A complete list is available in Table 1. The two genes with the strongest interactions are Growth Factor Receptor-bound Protein 2 (GRB2; $\kappa^{\alpha}_{\max} = 2.04$), which has a stronger association in the ER- group, and Adenomatous Polyposis Coli (APC; $\kappa^{\alpha}_{\max} = 1.91$), which has a stronger association in the ER+ group. Both genes have been hypothesized to be associated with breast cancer carcinogenesis (Daly et al., 1994; Jin et al., 2001).

2.6 Discussion

We have proposed a general framework for inference about absence/presence qualitative interactions. We argued that naïve procedures rely upon untenable conditions because the absence/presence hypothesis is ill-posed. We thus proposed to relax the problem in order to conduct well-calibrated inference that maintains the absence/presence interpretation and only requires mild assumptions.

In simulations, we found that our methodology has low power when signal is weak or sample sizes are small. To an extent, this is just a feature of the problem; naturally, one would require even more information to detect qualitative interactions than what is required to detect quantitative interactions. However, we provide no guarantee that our methodology is optimal, as tests for composite hypotheses based upon supremum p-values can be

Gene	ER+ Log HR (SE)	ER- Log HR (SE)	$\kappa^{\alpha}_{\max}$
GRB2	-0.06 (0.31)	-1.66 (0.68)	2.04
APC	1.34 (0.32)	-0.09 (0.33)	1.91
BAX	-1.05 (0.24)	0.04 (0.36)	1.53
PIK3CA	1.13 (0.28)	0.14 (0.32)	1.51
SOS2	1.13 (0.36)	-0.1 (0.37)	1.33
MAP2K2	-0.87 (0.27)	0.03 (0.35)	1.22
GADD45G	-0.52 (0.13)	-0.07 (0.19)	1.21
HES5	0.02 (0.2)	0.51 (0.18)	1.19
WNT2	-0.36 (0.09)	0 (0.17)	1.14
DLL4	0.09 (0.2)	0.68 (0.27)	1.10
FRAT2	-1.22 (0.31)	-0.45 (0.29)	1.08
SOS1	1.19 (0.3)	-0.34 (0.42)	1.01

Table 2.1: KEGG genes that are more strongly associated with time to death in one ER group than the other, i.e., $\kappa^{\alpha}_{\max} > 1$ with $\alpha = .10$.

conservative in practice (Bayarri and Berger, 2000).

Nonetheless, our framework is interpretable and provides a natural approach for quantifying differences in strength of association by sub-population in general settings. Though we only considered measures of marginal association in our examples, our method can be used with conditional measures of association as well; we only require that asymptotically normal estimates are available. In particular, our approach remains valid in the high-dimensional setting, where asymptotically normal estimates can be obtained using the techniques of, e.g., van de Geer et al. (2014) and Zhang and Zhang (2014). Finally, our method can be useful in the analysis of genomics data, as we demonstrated in our example.

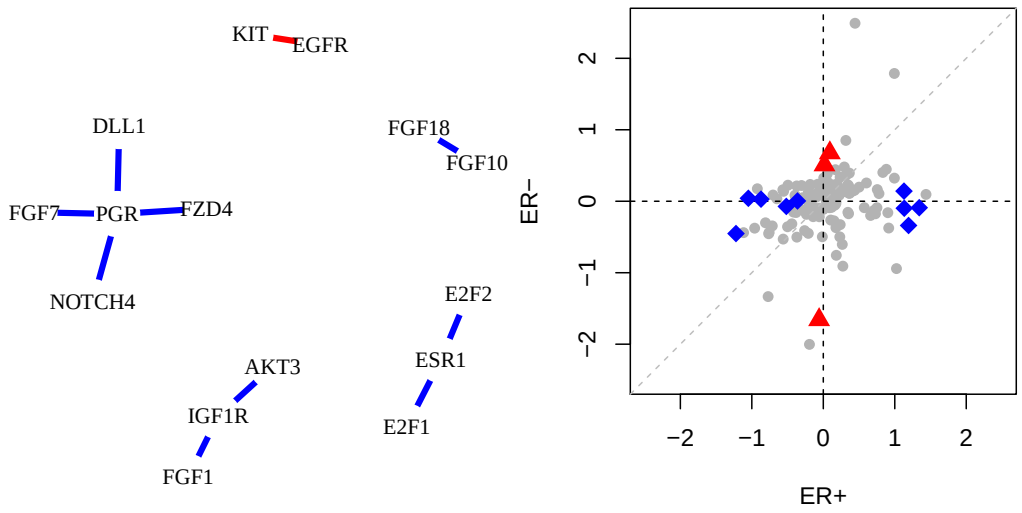


Figure 2.7: (Left) Pairs of genes in the KEGG breast cancer pathway for which we reject the relative difference hypothesis with $\kappa = 2$. Blue edges indicate associations that are stronger in the ER+ group, and red edges indicate associations that are stronger in the ER- group. (Right) Log hazard ratios for KEGG genes in ER+ and ER- groups. The gray dashed line represents the 45-degree line. Blue diamonds and red triangles indicate genes for which $\kappa^{\alpha}_{\max} > 1$ with $\alpha = .10$, where the largest log hazard ratio is in the ER+ group and ER- group, respectively.

Chapter 3

COVARIATE-ADJUSTED INFERENCE FOR DIFFERENTIAL ANALYSIS OF HIGH-DIMENSIONAL NETWORKS

3.1 Introduction

Complex diseases are often associated with aberrations in biological networks, such as gene regulatory networks and brain functional or structural connectivity networks (Barabási et al., 2011). Performing differential network analysis, or identifying connections in biological networks that affect disease condition, can provide insights into the disease mechanisms and lead to the identification of network-based biomarkers (Ideker and Krogan, 2012; de la Fuente, 2010).

Probabilistic graphical models are commonly used to summarize the conditional independence structure of a set of nodes in a biological network. A common approach to differential network analysis is to first estimate the graph corresponding to each disease condition and then assess between-condition differences in the graph. For instance, when using Gaussian graphical models, one can learn the network by estimating the inverse covariance matrix using the graphical LASSO (Friedman et al., 2008); one can then identify changes in the inverse covariance matrix associated with disease condition (Zhao et al., 2014; Xia et al., 2015; He et al., 2019). Alternatively, the condition-specific networks can be estimated using neighborhood selection (Meinshausen and Bühlmann, 2006); in this approach, partial correlations among nodes are estimated by fitting a series of linear regressions in which one node is treated as the outcome, and the remaining nodes are treated as regressors. Changes in the network can then be delineated from differences in the regression coefficients by disease condition (Belilovsky et al., 2016; Xia et al., 2018). More generally, the condition-specific networks are often modeled using exponential family models with pairwise interaction terms

(Lin et al., 2016; Yang et al., 2015; Yu et al., 2019, 2020).

In some instances, these approaches to differential network analysis may lead to the detection of between-group differences in biological networks that do not have a causal interpretation. For example, this can occur when the condition-specific networks depend on covariates (e.g., age and sex) because between-group network differences can be induced by *confounding variables*, i.e., variables that are associated with both the within-group networks, and the disease condition. In such cases, the network differences by disease condition may only reflect the association between the confounding variable and the disease. It is therefore important to account for the relationship between covariates and biological networks when performing differential network analysis. Adjustment for confounding variables in observational studies has been studied extensively in many contexts, e.g., in univariate outcome regression (McNamee, 2005). In this chapter, we use similar ideas to develop an approach for confounder adjustment in differential network analysis.

In this chapter, we propose a two-sample test for differential network analysis that accounts for differing covariate effects on the networks, within each group. More specifically, we propose to perform covariate-adjusted inference using a class of pairwise interaction models for the within-group networks. Our approach treats each group-specific network as a function of the covariates. It then performs a hypothesis test for differences between these functions. To accommodate the high-dimensional setting, in which the number of nodes in the network is large relative to the number of samples collected, we propose to estimate the networks using a regularized estimator and to perform hypothesis testing using a bias-corrected version of the regularized estimate (van de Geer, 2016).

Our proposal is related to existing literature on modeling networks as functions of a small number of variables. For example, there are various proposals for estimating high-dimensional inverse covariance matrices, conditional upon continuous low-dimensional features (Zhou et al., 2010; Wang and Kolar, 2014). Also related are methods for regularized estimation of high-dimensional varying coefficient models, wherein the regression coefficients are functions of a small number of covariates (Wang and Xia, 2009). Our method is similar

but places a particular emphasis on hypothesis testing in order to assess the statistical significance of observed differences in the networks. Our approach is the first, to the best of our knowledge, to perform covariate-adjusted hypothesis tests for differential network analysis.

The rest of the chapter is organized as follows. In Section 3.2, we begin with a broad overview of our proposed framework for covariate-adjusted differential network analysis in pairwise interaction exponential family models and introduce some working examples. In the following sections, we specialize our framework by considering two different approaches for estimation and inference: In Section 3.3, we describe a method that uses neighborhood selection (Meinshausen and Bühlmann, 2006; Chen et al., 2015; Yang et al., 2015), and in Section 3.4, we discuss an alternative estimation approach that utilizes the score matching framework of Hyvärinen (2005, 2007). We assess the performance of our proposed methodology on synthetic data in Section 3.5 and apply it to a breast cancer data set from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) in Section 3.6. We conclude with a brief discussion in Section 3.7.

3.2 Overview of the Proposed Framework

3.2.1 Differential network analysis without covariate adjustment

To formalize our problem, we begin by introducing some notation. We compare networks between two groups, labeled by $g \in \{\mathrm{I}, \mathrm{II}\}$. We obtain measurements of p variables $X^g = \left(X_1^g, \ldots, X_p^g\right)^\top$, corresponding to nodes in a graphical model (Maathuis et al., 2018), on n^{I} subjects in group I and n^{II} subjects in group II. We define $\mathcal{X} \subseteq \mathbb{R}^p$ as the sample space of X^g. Let $X_{i,j}^g$ denote the data for node j for subject i in group g, and let $\mathbf{X}_j^g = (X_{1,j}^g, \ldots, X_{n^g,j}^g)^\top$ be an n^g-dimensional vector of measurements on node j for group g.

Our objective is to determine whether the association between variables X_j and X_k, conditional upon all other variables, differs by group. Our approach is to specify a model for X^g in which the conditional dependence between any two nodes X_j^g and X_k^g is represented by a single scalar parameter $\beta_{j,k}^{g,*}$. If the association between nodes j and k is the same in both

groups I and II, $\beta_{j,k}^{\mathrm{I},*} = \beta_{j,k}^{\mathrm{II},*}$. Conversely, if $\beta_{j,k}^{\mathrm{I},*} \neq \beta_{j,k}^{\mathrm{II},*}$, we say nodes j and k are *differentially connected*. We assess for differential connectivity by performing a test of the null hypothesis

$$H_{j,k}^0 : \beta_{j,k}^{\mathrm{I},*} = \beta_{j,k}^{\mathrm{II},*}. \tag{3.1}$$

We consider a general class of exponential family pairwise interaction models. For $x = (x_1, \ldots, x_p)^\top$, we assume the density function for X^g takes the form

$$f^{g,*}(x) = \exp\left(\sum_{j=1}^{p} \mu_j(x_j) + \sum_{j=1}^{p}\sum_{k=1}^{j} \beta_{j,k}^{g,*}\psi_{j,k}(x_j, x_k) - U\left(\boldsymbol{\beta}^{g,*}\right)\right), \tag{3.2}$$

where $\psi_{j,k}$ and μ_j are fixed and known functions, $\boldsymbol{\beta}^{g,*}$ is a $p \times p$ matrix with elements $\beta_{j,k}^{g,*}$, and $U(\boldsymbol{\beta}^{g,*})$ is the log-partition function. The dependence between X_j^g and X_k^g is measured by $\beta_{j,k}^{g,*}$, and nodes j and k are conditionally independent in group g if and only if $\beta_{j,k}^{g,*} = 0$.

This class of exponential family distributions is rich and includes several models that have been studied previously in the graphical modeling literature. One such example is the Gaussian graphical model, perhaps the most widely-used graphical model for continuous data. For $x \in \mathbb{R}^p$ the density function for mean-centered Gaussian random vectors can be expressed as

$$f^{g,*}(x) \propto \exp\left(-\sum_{j=1}^{p}\sum_{k=1}^{j} \beta_{j,k}^{g,*} x_j x_k\right), \tag{3.3}$$

and is thus a special case of (3.2) with $\psi_{j,k} = -x_j x_k$ and $\mu_j = 0$. The non-negative Gaussian density, which takes the form of (3.3) with the constraint that x takes values in $\mathbb{R}_+^p$, also belongs to the exponential family class. Another canonical example is the Ising model, commonly used for studying conditional dependencies among binary random variables. For $x \in \{0, 1\}^p$, the density function for the Ising model can be expressed as

$$f^{g,*}(x) \propto \exp\left(\sum_{j=1}^{p}\sum_{k=1}^{j} \beta_{j,k}^{g,*} x_j x_k\right).$$

Additional examples include the Poisson model, the exponential graphical model, and conditionally-specified mixed graphical models (Yang et al., 2015; Chen et al., 2015).

When asymptotically normal estimates of $\beta_{j,k}^{\mathrm{I},*}$ and $\beta_{j,k}^{\mathrm{II},*}$ are available, one can perform a well-calibrated test of $H_{j,k}^0$ based on the difference between the estimates. In many cases, asymptotically normal estimates can be obtained using well-established methodology. For instance, when the log-partition function $U(\boldsymbol{\beta}^{g,*})$ is available in closed form and is tractable, one can obtain estimates via (penalized) maximum likelihood. This is a standard approach in the Gaussian setting, in which case the log-partition function is easy to compute. However, this is not the case for other exponential family models. Likelihood-based estimation strategies are thus generally difficult to implement. In this chapter, we consider two alternative strategies that have been proposed to overcome these computational challenges and are more broadly applicable.

The first approach we discuss is neighborhood selection (Chen et al., 2015; Meinshausen and Bühlmann, 2006; Yang et al., 2015). Consider a sub-class of exponential family graphical models for which the conditional density function for any node X_j^g given the remaining nodes belongs to a univariate exponential family model. Because the log-partition function in univariate exponential family models is available in closed form, it is computationally feasible to estimate each conditional density function. By estimating the conditional density functions, one can identify the *neighbors* of nodes j, that is, the nodes upon which the conditional distribution depends. This approach was first proposed as an alternative to maximum likelihood estimation for estimating Gaussian graphical models (Meinshausen and Bühlmann, 2006). To describe our approach, we focus on the Gaussian case, though this approach is more widely applicable and can be used for modeling dependencies among, e.g., Poisson, binomial, and exponential random variables as well (Chen et al., 2015; Yang et al., 2015).

In Gaussian graphical models, the dependency of node j on all other nodes can be determined based on the linear model

$$\mathbb{E}\left[X_j^g | X_1^g, \ldots, X_p^g\right] = \beta_{j,0}^{g,*} + \sum_{k \neq j} \beta_{j,k}^{g,*} X_k^g. \tag{3.4}$$

The regression coefficients $\beta_{j,k}^{g,*}$ measure the strength of linear association between nodes j

and k conditional upon all other nodes and are zero if and only if nodes j and k are conditionally independent; $\beta_{j,0}^{g,*}$ is an intercept term and is zero if all nodes are mean-centered. (We acknowledge a slight abuse of notation here, as the regression coefficients in (3.4) are not equivalent to parameters in (3.2). However, either estimand fully characterizes conditional independence.) The network can thus be estimated by fitting a series of p linear regression models where one node is treated as the response, and all remaining nodes are treated as predictors. In the low-dimensional setting (i.e., $p \ll n^g$), statistically efficient and asymptotically normal estimates of the regression coefficients can be readily obtained via ordinary least squares. In high-dimensions (i.e., $p \geq n^g$), the ordinary least squares estimates are not well-defined, so to obtain consistent estimates we typically rely upon regularized estimators such as the LASSO and the elastic net (Tibshirani, 1996; Zou and Hastie, 2005). While regularized estimators are generally biased and have intractable sampling distributions, several methods have recently emerged for obtaining asymptotically normal estimates by correcting the bias of regularized estimators (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014).

The second computationally efficient approach we consider is to estimate the density function using the score matching framework of Hyvärinen (Hyvärinen, 2005, 2007). Hyvärinen derives a loss function for estimation of density functions for continuous random variables that is based on the gradient of the log-density with respect to the observations. As such, the score matching loss does not depend on the log-partition function in exponential family models. Moreover, when the joint distribution for X^g belongs to an exponential family model, the loss is quadratic in the unknown parameters, allowing for efficient computation. In low dimensions, the minimizer of the score matching loss is consistent and asymptotically normal. In high dimensions, one can obtain asymptotically normal estimates by minimizing a regularized version of the score matching loss to obtain an initial estimate (Lin et al., 2016; Yu et al., 2019) and subsequently correcting for the bias induced by regularization (Yu et al., 2020).

3.2.2 Covariate-adjusted differential network analysis

We now consider settings in which the within-group networks depend on covariates. We denote by W^g a q-dimensional random vector of covariate measurements for group g, and we define $\mathcal{W}$ as the sample space of W^g. Let $W^g_{i,r}$ refer to the value of covariate r for subject i in group g, and let $W^g_i = (W^g_{i,1}, \ldots, W^g_{i,q})^\top$ be a q-dimensional vector containing all covariates for subject i in group g. We assume the number of covariates is small relative to the sample size (i.e., $q \ll n^g$).

To study the dependence of the within-group networks on the covariates, we specify a model for the nodes X^g given the covariates W^g that allows the inter-node dependencies to vary as a function of W^g. The model defines a function $\eta^g_{j,k} : \mathcal{W} \to \mathbb{R}$ that takes as input a vector of covariates and returns a measure of association between nodes j and k for a subject in group g with identical covariates. One can interpret $\eta^{g,*}_{j,k}$ as a conditional version of $\beta^{g,*}_{j,k}$ in (3.2), given the covariates.

We assume that $\eta^{g,*}_{j,k}$ can be written as a low-dimensional linear basis expansion in W^g of dimension d — that is,

$$\eta^{g,*}_{j,k}\left(W^g\right) = \left\langle \phi\left(W^g\right), \alpha^{g,*}_{j,k} \right\rangle,$$

where $\phi : \mathbb{R}^q \to \mathbb{R}^d$ is a map from a set of covariates to its expansion, $\alpha^{g,*}_{j,k}$ is a d-dimensional vector, and $\langle \cdot, \cdot \rangle$ denotes the vector inner product. Let $\phi_c(w)$ refer to the c-th element of $\phi(w)$. One can take the simple approach of specifying ϕ as a linear basis, $\phi(w) = (1, w_1, \ldots, w_q)$ for $w \in \mathbb{R}^q$, though more flexible choices such as polynomial or B-spline bases can also be considered. It may be preferable to specify ϕ so that $\eta^{g,*}_{j,k}$ is an additive function of the covariates, as this allows one to easily assess the effect of any specific covariate on the network, by estimating the sub-vector of $\alpha^{g,*}_{j,k}$ that is relevant to the covariate of interest.

When the association between nodes j and k does not depend on group membership, $\eta^{\mathrm{I},*}_{j,k}(w) = \eta^{\mathrm{II},*}_{j,k}(w)$ for all w, and $\alpha^{\mathrm{I},*}_{j,k} = \alpha^{\mathrm{II},*}_{j,k}$. In other words, if one subject from group I and another subject from group II have identically-valued covariates, the corresponding measure of association between nodes j and k is also the same. In the covariate-adjusted

setting, we say that nodes j and k are differentially connected if there exists w such that $\eta_{j,k}^{\mathrm{I},*}(w) \neq \eta_{j,k}^{\mathrm{II},*}(w)$, or equivalently, if $\alpha_{j,k}^{\mathrm{I},*} \neq \alpha_{j,k}^{\mathrm{II},*}$. We can thus assess differential connectivity between nodes j and k by testing the null hypothesis

$$G_{j,k}^0 : \alpha_{j,k}^{\mathrm{I},*} = \alpha_{j,k}^{\mathrm{II},*}. \tag{3.5}$$

Similar to the unadjusted setting, when asymptotically normal estimates of $\alpha_{j,k}^{\mathrm{I},*}$ and $\alpha_{j,k}^{\mathrm{II},*}$ are available, a calibrated test can be constructed based on the difference between the estimates.

We now specify a form for the conditional distribution of X^g given W^g as a generalization of the exponential family pairwise interaction model (3.2). We assume the conditional density for X^g given W^g can be expressed as

$$f^{g,*}(x|w) \propto \exp\left(\sum_{j=1}^{p} \mu_j(x_j) + \sum_{j=1}^{p}\sum_{k=1}^{j} \eta_{j,k}^{g,*}(w)\psi_{j,k}(x_j, x_k) + \sum_{j=1}^{p}\sum_{c=1}^{d} \theta_{j,c}^{g,*}\zeta_{j,c}\left(x_j, \phi_c(w)\right)\right), \tag{3.6}$$

where $w = (w_1, \ldots, w_q)^\top$, and the proportionality is up to a normalizing constant that does not depend on x. Above, $\zeta_{j,c}$ is a fixed and known function, and the main effects of the covariates on X^g are represented by the scalar parameters $\theta_{j,c}^{g,*}$. The conditional dependence between nodes j and k, given all other nodes and given that $W^g = w$ is quantified by $\eta_{j,k}^{g,*}(w)$, and $\eta_{j,k}^{g,*}(w) = 0$ if and only if nodes j and k are conditionally independent at w. One can thus view $\eta_{j,k}^{g,*}$ as a conditional version of $\beta_{j,k}^{g,*}$ in (3.6).

Either of the estimation strategies introduced in Section 3.2.1 can be used to perform covariate-adjusted inference. When the conditional distribution of each node given the remaining nodes *and the covariates* belongs to a univariate exponential family model, the covariate-dependent network can be estimated using neighborhood selection because the node conditional distributions can be estimated efficiently with likelihood-based methods. Alternatively, we can estimate the conditional density function (3.6) using score matching.

As a working example, we again consider estimation of covariate-dependent Gaussian networks using neighborhood selection. Suppose the conditional distribution of X^g given

W^g takes the form

$$f^{g,*}(x|w) \propto \exp\left(-\sum_{j=1}^{p}\sum_{k=1}^{j}\eta_{j,k}^{g,*}(w)x_j x_k - \sum_{j=1}^{p}\sum_{c=1}^{d}\theta_{j,c}^{g,*}x_j\phi_c(w)\right). \tag{3.7}$$

Then the dependencies of node j on all other nodes can be determined based on the following varying coefficient model (Hastie and Tibshirani, 1993):

$$\mathbb{E}\left[X_j^g|X_1^g,\ldots,X_p^g,W^g\right] = \eta_{j,0}^{g,*}(W^g) + \sum_{k\neq j}\eta_{j,k}^{g,*}\left(W^g\right)X_k^g. \tag{3.8}$$

The varying coefficient model is a generalization of the linear model that treats the regression coefficients as functions of the covariates. In (3.8), $\eta_{j,k}^{g,*}(w)$ returns a regression coefficient that quantifies the linear relationship between nodes j and k for subjects in group g with covariates equal to w. Then X_j^g and X_k^g are conditionally independent given all other nodes and given $W^g = w$ if and only if $\eta_{j,k}^{g,*}(w) = 0$. The varying coefficients $\eta_{j,k}^{g,*}$ can thus be viewed as a conditional version of the regression coefficients in (3.4). (We have again abused the notation, as the varying coefficient functions in (3.8) are not equal to the parameters in (3.7), though both functions are zero for the same values of w). The intercept term $\eta_{j,0}^{g,*}$ accounts for the main effect of W^g on X_j^g. We can remove this main effect term by first centering the nodes X_j^g about their conditional mean given W^g (which can be estimated by performing a linear regression of X_j^g on $\phi(W^g)$).

In Sections 3.3 and 3.4, we discuss construction of asymptotically normal estimators of $\alpha_{j,k}^{g,*}$ in the low- and high-dimensional settings using neighborhood selection and score matching. Before proceeding, we first examine the connection between the null hypotheses $H_{j,k}^0$ and $G_{j,k}^0$.

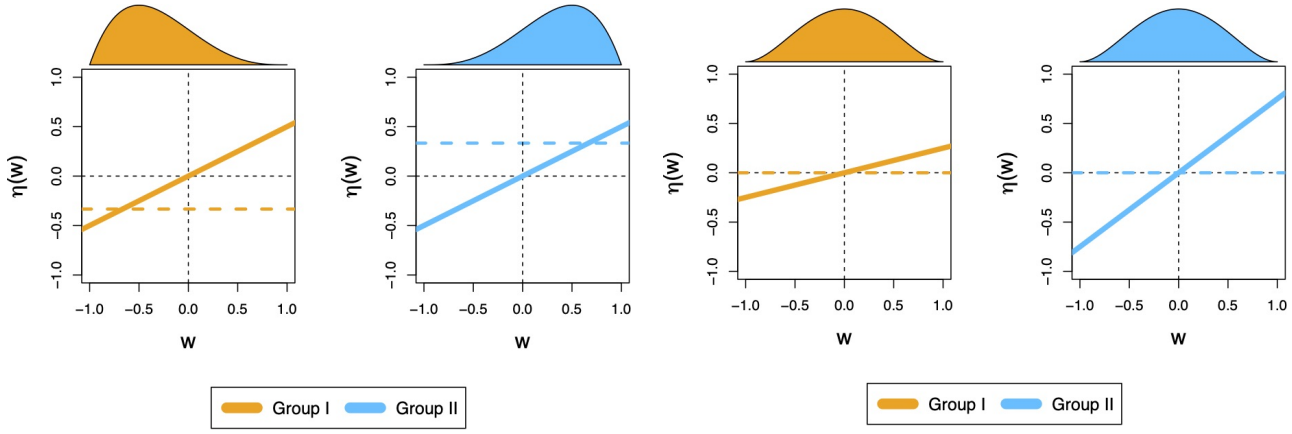
3.2.3 The relationship between hypotheses $H_{j,k}^0$ and $G_{j,k}^0$

As there is generally no equivalence between marginal and conditional associations, it is possible that $\beta_{j,k}^{g,*}$ in (3.2) differs with group while $\eta_{j,k}^{g,*}$ in (3.6) does not, and vice versa. However, while hypotheses $H_{j,k}^0$ in (3.1) and $G_{j,k}^0$ in (3.5) are not equivalent, they are still related. We illustrate this by providing an example below. Suppose we are using neighborhood

selection to perform differential network analysis in the Gaussian setting, so we are making a comparison of linear regression coefficients between the two groups. Suppose further that the within-group networks depend on single scalar covariate W^g, and the nodes are centered about their conditional mean given W^g. One can show that the regression coefficients $\beta_{j,k}^{g,*}$ are equal to the average of their conditional versions $\eta_{j,k}^{g,*}(W^g)$. That is, $\beta_{j,k}^{g,*} = \mathbb{E}[\eta_{j,k}^{g,*}(W^g)]$. Now, suppose $G_{j,k}^0$ holds. If W^{I} and W^{II} do not share the same distribution (e.g., the covariate tends to take higher values in group I than in group II), the average conditional inter-node association may differ, and $H_{j,k}^0$ may not hold. Although the conditional association between nodes, given the covariate, does not differ by group, the *average* conditional association does differ, as illustrated in Figure 3.1a. In such a scenario, the difference in the average conditional association is induced by the dependence of the covariate on group membership and the dependence of the inter-node association on the covariate. Thus, inequality of $\beta_{j,k}^{\mathrm{I},*}$ and $\beta_{j,k}^{\mathrm{II},*}$ does not necessarily capture a meaningful association between the network and group membership. Similarly when $H_{j,k}^0$ holds, it is possible that $\eta_{j,k}^{\mathrm{I},*} \neq \eta_{j,k}^{\mathrm{II},*}$. For instance, suppose that the distribution of the covariate is the same in both groups, and $\mathbb{E}[\eta^g(W^g)] = 0$ in both groups. If the between-node association depends more strongly upon the covariates in one group than the other, $G_{j,k}^0$ will be false. This example is depicted in Figure 3.1b. In this scenario, adjusting for covariates should provide improved power to detect differential connections. We note that for other distributions, it does not necessarily hold that $\beta_{j,k}^{g,*} = \mathbb{E}[\eta_{j,k}^{g,*}(W^g)]$, but regardless, there is generally no equivalence between hypotheses $H_{j,k}^0$ and $G_{j,k}^0$.

3.3 Covariate-adjusted differential network analysis using neighborhood selection

In this section, we describe in detail an approach for covariate-adjusted differential network analysis using neighborhood selection. To simplify our presentation, we focus on Gaussian graphical models, though this strategy is generally applicable to graphical models for which the node conditional distributions belong to univariate exponential family models.



(a) (b)

Figure 3.1: Displayed are the association between nodes j and k, $\eta^g_{j,k}(\cdot)$, as a function of covariate W^g and the distribution of W^g in groups I and II. The average inter-node association is represented by the dashed colored lines. In (a), the average inter-node association depends on group membership, though the inter-node association given the covariate does not. In (b), the average inter-node association does not depend on group membership, though the conditional association between nodes given the covariate does depend on group membership.

3.3.1 Covariate adjustment via neighborhood selection in low dimensions

We first discuss testing the unadjusted null hypothesis $H^0_{j,k}$ in (3.1), where the $\beta^{g,*}_{j,k}$ are the regression coefficients in (3.4). Suppose, for now, that we are in the low-dimensional setting, so the number of nodes p is smaller than the sample sizes n^g, $g \in \{\mathrm{I}, \mathrm{II}\}$.

It is well-known that the regression coefficients can be characterized as the minimizers of the expected least squares loss — that is,

$$\boldsymbol{\beta}^{g,*}_j = (\beta^{g,*}_{j,1}, \ldots, \beta^{g,*}_{j,p})^\top = \underset{\beta_1,\ldots,\beta_p \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\left(X^g_j - \sum_{k \neq j} X^g_k \beta_k\right)^2\right].$$

One can obtain an estimate $\hat{\boldsymbol{\beta}}^g_j = (\hat{\beta}^g_{j,1}, \ldots, \hat{\beta}^g_{j,p})$ of $\boldsymbol{\beta}^{g,*}_j = (\beta^{g,*}_{j,1} \ldots, \beta^{g,*}_{j,p})$ by minimizing the

empirical average of the least squares, taking

$$\hat{\boldsymbol{\beta}}_j^g = \underset{\beta_1,\ldots,\beta_p\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n^g} \left\| \mathbf{X}_j^g - \sum_{k\neq j} \mathbf{X}_k^g \beta_k \right\|_2^2,$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. The ordinary least squares estimate $\hat{\boldsymbol{\beta}}_j^g$ is available in closed form and is easy to compute. The estimates $\hat{\beta}_{j,k}^g$ are unbiased, and, under mild assumptions, are approximately normally distributed for sufficiently large n^g — that is,

$$\hat{\beta}_{j,k}^g \sim N\left(\beta_{j,k}^{g,*}, \tau_{j,k}^g\right),$$

with $\tau_{j,k}^g > 0$ (though $\tau_{j,k}^g$ can be calculated in closed form, we omit the expression for brevity).

We construct a test of $H_{j,k}^0$ based on the difference between the estimates of the group-specific regression coefficients, $\hat{\beta}_{j,k}^{\mathrm{I}} - \hat{\beta}_{j,k}^{\mathrm{II}}$. When $H_{j,k}^0$ holds, $\hat{\beta}_{j,k}^{\mathrm{I}} - \hat{\beta}_{j,k}^{\mathrm{II}}$ is normally distributed with mean zero and variance $\tau_{j,k}^{\mathrm{I}} + \tau_{j,k}^{\mathrm{II}}$. Given a consistent estimate $\hat{\tau}_{j,k}^g$ of the variance, we can use the test statistic

$$T_{j,k} = \frac{\left(\hat{\beta}_{j,k}^{\mathrm{I}} - \hat{\beta}_{j,k}^{\mathrm{II}}\right)^2}{\hat{\tau}_{j,k}^{\mathrm{I}} + \hat{\tau}_{j,k}^{\mathrm{II}}},$$

which follows a chi-square distribution with one degree of freedom under the null for n^{I} and n^{II} sufficiently large. A p-value for $H_{j,k}^0$ can be calculated as

$$\rho_{j,k} = \mathbb{P}\left(\chi_1^2 > T_{j,k}\right).$$

In the low-dimensional setting, performing a covariate-adjusted test is similar to performing the unadjusted test. We can obtain an estimate $\hat{\boldsymbol{\alpha}}_j^g = \left((\hat{\alpha}_{j,1}^g)^\top, \ldots, (\hat{\alpha}_{j,p}^g)^\top\right)^\top$ of $\boldsymbol{\alpha}_j^{g,*} = \left((\alpha_{j,1}^{g,*})^\top, \ldots, (\alpha_{j,p}^{g,*})^\top\right)^\top$ by minimizing the empirical average of the least squares loss

$$\hat{\boldsymbol{\alpha}}_j^g = \underset{\alpha_{j,1},\ldots,\alpha_{j,p}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n^g} \sum_{i=1}^{n^g} \left(X_{i,j}^g - \sum_{k\neq j} \left\langle \phi\left(W_i^g\right), \alpha_{j,k} \right\rangle X_{i,k}^g \right)^2. \tag{3.9}$$

To simplify the presentation, we introduce additional notation that allows us to rewrite (3.9) in a condensed form. Let $\mathcal{V}_k^g$ be the $n^g \times d$ matrix

$$\mathcal{V}_k^g = \begin{pmatrix} X_{1,k}^g \times \phi\left(W_1^g\right) \\ \vdots \\ X_{n^g,k}^g \times \phi\left(W_{n^g}^g\right) \end{pmatrix}. \tag{3.10}$$

We can now equivalently express (3.9) as

$$\hat{\boldsymbol{\alpha}}_j^g = \underset{\alpha_{j,1},\ldots,\alpha_{j,p} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n^g} \left\| \mathbf{X}_j^g - \sum_{k \neq j} \mathcal{V}_k^g \alpha_{j,k} \right\|_2^2. \tag{3.11}$$

Again, $\hat{\alpha}_{j,k}^g$ is an unbiased and approximately normal for sufficiently large n^g, satisfying

$$\hat{\alpha}_{j,k}^g \sim N\left(\alpha_{j,k}^{g,*}, \Omega_{j,k}^g\right),$$

where $\Omega_{j,k}^g$ is a positive definite matrix of dimension $d \times d$ (though a closed form expression is available, we omit it here for brevity).

We construct a test of $G_{j,k}^0$ based on $\hat{\alpha}_{j,k}^{\mathrm{I}} - \hat{\alpha}_{j,k}^{\mathrm{II}}$. Under the null hypothesis, $\hat{\alpha}_{j,k}^{\mathrm{I}} - \hat{\alpha}_{j,k}^{\mathrm{II}}$ follows a normal distribution with mean zero and variance $\Omega_{j,k}^{\mathrm{I}} + \Omega_{j,k}^{\mathrm{II}}$. Given a consistent estimate $\hat{\Omega}_{j,k}^g$ of $\Omega_{j,k}^g$, we can test $G_{j,k}^0$ using the test statistic

$$S_{j,k} = \left(\hat{\alpha}_{j,k}^{\mathrm{I}} - \hat{\alpha}_{j,k}^{\mathrm{II}}\right)^\top \left(\hat{\Omega}_{j,k}^{\mathrm{I}} + \hat{\Omega}_{j,k}^{\mathrm{II}}\right)^{-1} \left(\hat{\alpha}_{j,k}^{\mathrm{I}} - \hat{\alpha}_{j,k}^{\mathrm{II}}\right).$$

Under the null, the test statistic follows a chi-squared distribution with d degrees of freedom, and a p-value can therefore be calculated as

$$\mathbb{P}\left(\chi_d^2 > S_{j,k}\right).$$

3.3.2 Covariate adjustment via neighborhood selection in high dimensions

The methods described in Section 3.3.1 are only appropriate when the number of nodes p is small relative to the sample size. Model (3.8) has $(p-1)d$ parameters, so the least squares estimator of Section 3.3.1 provides stable estimates as long as n^{I} and n^{II} are larger than

$(p-1)d$. However, in the high-dimensional setting, where the the number of parameters exceeds the sample size, the ordinary least squares estimates are not well-defined.

To fit the varying coefficient model (3.8) in the high-dimensional setting, we use a regularized estimator that relies upon an assumption of *sparsity* in the networks. The sparsity assumption requires that within each group only a small number of nodes are partially correlated, meaning that in (3.8), only a few of the vectors $\alpha_{j,k}^{g,*}$ are nonzero. To leverage the sparsity assumption, we propose to use the group LASSO estimator (Yuan and Lin, 2006):

$$\tilde{\boldsymbol{\alpha}}_j^g = \underset{\alpha_{j,1},\ldots,\alpha_{j,p}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n^g} \left\| \mathbf{X}_j^g - \sum_{k\neq j} \mathcal{V}_k^g \alpha_{j,k} \right\|_2^2 + \lambda \sum_{k\neq j} \|\alpha_{j,k}\|_2 , \tag{3.12}$$

where $\lambda > 0$ is a tuning parameter. The group LASSO provides a sparse estimate and sets some $\tilde{\alpha}_{j,k}$ to be exactly zero, resulting in networks with few edges. The level of sparsity of $\tilde{\boldsymbol{\alpha}}_j^g$ is determined by λ, with higher λ values forcing more $\tilde{\alpha}_{j,k}$ to zero. We discuss selection of the tuning parameter in Section 3.5.1.

Though the group LASSO provides a consistent estimate of $\boldsymbol{\alpha}_j^{g,*}$, the estimate is *not* approximately normally distributed. The group LASSO estimate of $\alpha_{j,k}^{g,*}$ retains a bias that diminishes at the same rate as the standard error. As a result, the group LASSO estimator has a non-standard sampling distribution that cannot be derived analytically. However, we can obtain approximately normal estimates of $\alpha_{j,k}^{g,*}$ by correcting the bias of $\tilde{\alpha}_{j,k}^g$, as was first proposed to obtain normal estimates for the classical ℓ_1-penalized version of the LASSO (van de Geer et al., 2014; Zhang and Zhang, 2014). These "de-biased" or "de-sparsified" estimators can been shown to be approximately normal with moderately large samples even in the high-dimensional setting; they are therefore suitable for hypothesis testing. Our approach is to use a de-biased version of the group LASSO. Bias correction in group LASSO problems is well-studied (van de Geer, 2016; Honda, 2019; Mitra and Zhang, 2016), so we are able to perform covariate-adjusted inference by applying previously-developed methods.

The bias of the group LASSO estimate can be written as

$$\delta_{j,k}^g = \mathbb{E}\left[\tilde{\alpha}_{j,k}^g\right] - \alpha_{j,k}^{g,*},$$

where $\delta^g_{j,k}$ is a nonzero d-dimensional vector (recall d is the dimension of $\alpha^{g,*}_{j,k}$). Our approach is to obtain an estimate of the bias $\tilde{\delta}_{j,k}$ and to use a de-biased estimator, defined as

$$\check{\alpha}^g_{j,k} = \tilde{\alpha}^g_{j,k} - \tilde{\delta}^g_{j,k}.$$

For a suitable choice of $\tilde{\delta}_{j,k}$, the bias-corrected estimator is approximately normal for a sufficiently large sample size n^g under mild conditions, i.e.,

$$\check{\alpha}^g_{j,k} \sim N\left(\alpha^{g,*}_{j,k}, \Omega^g_{j,k}\right),$$

where the variance $\Omega^g_{j,k}$ is a positive definite matrix, for which we obtain an estimate $\check{\Omega}^g_{j,k}$. We provide a derivation for the bias-correction and the form of our variance estimate in Appendix B.1.

Similar to Section 3.3.1, we test the null hypothesis $G^0_{j,k}$ in (3.5) using the test statistic

$$S_{j,k} = \left(\check{\alpha}^{\mathrm{I}}_{j,k} - \check{\alpha}^{\mathrm{II}}_{j,k}\right)^\top \left(\check{\Omega}^{\mathrm{I}}_{j,k} + \check{\Omega}^{\mathrm{II}}_{j,k}\right)^{-1} \left(\check{\alpha}^{\mathrm{I}}_{j,k} - \check{\alpha}^{\mathrm{II}}_{j,k}\right).$$

The test statistic asymptotically follows a chi-squared distribution with d degrees of freedom under the null hypothesis.

3.4 Covariate-adjusted differential network analysis using score matching

In this section, we discuss covariate-adjustment using the score matching framework introduced in Section 3.2. We first describe the score matching estimator in greater detail and then specialize the framework to estimation of pairwise exponential family graphical models in the low- and high dimensional settings. As shown later in this section, for exponential family distributions with continuous support, the score matching loss function is a quadratic function of parameters, providing a computationally-efficient framework for estimating graphical models.

3.4.1 The score matching framework

We begin by providing a brief summary of the score matching framework (Hyvärinen, 2005, 2007). Let $Z \in \mathcal{Z} \subseteq \mathbb{R}^p$ be a random vector generated from a distribution with density

function h^*. For any candidate density h, we denote the gradient and Laplician of the log-density by

$$\nabla \log h(z) = \left\{ \frac{\partial}{\partial z_j} \log h(z) \right\} \in \mathbb{R}^p; \qquad \Delta \log h(z) = \sum_{j=1}^{p} \frac{\partial^2}{\partial z_j^2} \log h(z_j).$$

The *score matching loss* L is defined as a measure of divergence between a candidate density function h and the true density h^*:

$$L(h) = \int \|\nabla \log h(z) - \nabla \log h^*(z)\|_2^2 \, h^*(z) dz = \mathbb{E}\left[\|\nabla \log h(Z) - \nabla \log h^*(Z)\|_2^2 \right]. \quad (3.13)$$

It is apparent that the score matching loss is minimized when $h = h^*$. A natural approach to constructing an estimator for h^* would then be to minimize the empirical score matching loss given observations $Z_1, \ldots, Z_n$, defined as

$$L_n(h) = \frac{1}{n} \sum_{i=1}^{n} \|\nabla \log h\,(Z_i) - \nabla \log h^*\,(Z_i)\|_2^2 .$$

Because the score matching loss function takes as input the gradient of the log density function, the loss does not depend on the normalizing constant. This makes score matching appealing when the normalizing constant is intractable.

The empirical loss seemingly depends on prior knowledge of h^*. However, if $h(z)$ and $\|h(z)\|_2$ both tend to zero as z approaches the boundary of $\mathcal{Z}$, a partial integration argument can be used to show that the score matching loss can be expressed as

$$L(h) = \int \left\{ \Delta \log h(z) + \frac{1}{2} \|\nabla \log h(z)\|_2^2 \right\} h^*(z) dz + \text{const.}, \quad (3.14)$$

where 'const.' is a term that does not depend on h. We can therefore estimate h^* by minimizing an empirical version of the score matching loss that does not depend on h^*. We can express the empirical loss as

$$L_n(h) = \frac{1}{n} \sum_{i=1}^{n} \Delta \log h(Z_i) + \frac{1}{2} \|\nabla \log h(Z_i)\|_2^2 .$$

The score matching loss is particularly appealing for exponential family distributions with continuous support, as it leads to a quadratic optimization function (Lin et al., 2016).

However, when Z is non-negative, the arguments used to express (3.13) as (3.14) fail because $h(z)$ and $\|\nabla h(z)\|_2$ do not approach zero at the boundary. We can overcome this problem by instead considering the *generalized score matching framework* (Yu et al., 2019; Hyvärinen, 2007) as an extension that is suitable for non-negative data. Let $v_1, \ldots, v_p : \mathbb{R}^+ \to \mathbb{R}^+$ be positive and differentiable functions, let $v(z) = (v_1(z_1), \ldots, v_p(z_p))^\top$, let $\dot{v}_j$ denote the derivative of v_j, and let $\circ$ denote the element-wise product operator. The generalized score matching loss is defined as

$$L(h) = \int \left\| \{\nabla \log h(z) - \nabla \log h^*(z)\} \circ v^{1/2}(z) \right\|_2^2 h^*(z) dz, \tag{3.15}$$

and is also minimized when $h = h^*$. As for the original score matching loss (3.13), the generalized score matching loss seemingly depends on prior knowledge of h^*. However, under mild technical conditions on h and v (see Appendix B.2.1), the loss in (3.15) can be rewritten as

$$L(h) = \int \left[\sum_{j=1}^{p} \dot{v}_j(z_j) \left\{ \frac{\partial \log h(z_j)}{\partial z_j} \right\} + v_j(z_j) \left\{ \frac{\partial^2 \log h(z)}{\partial z_j^2} \right\} + \frac{1}{2} v_j(z_j) \left\{ \frac{\partial \log h(z)}{\partial z_j} \right\}^2 \right] h^*(z) dz. \tag{3.16}$$

The generalized score matching loss thus no longer depends on h^*, and an estimator can be constructed by minimizing the empirical version of (3.16) with respect to h. To this end, the original generalized score matching estimator considered $v_j(z_j) = z_j^2$ (Hyvärinen, 2007). In this case, it becomes necessary to estimate high moments of h^*, leading to poor performance of the estimator. It has been shown that by instead taking v as a slowly increasing function, such as $v_j(z_j) = \log(1 + v_j)$, one obtains improved theoretical results and better empirical performance (Yu et al., 2019).

3.4.2 Covariate adjustment in high-dimensional exponential family models via score matching

In this sub-section, we discuss construction of asymptotically normal estimators for the parameters of the exponential family pairwise interaction model (3.6) using the generalized

score matching framework. To simplify our presentation, we consider the setting in which we are only interested in studying the connectedness between one node X_j^g and all other neighboring nodes in the network. To this end, it suffices to estimate the conditional density of X_j^g given all other nodes and the covariates W^g. A similar approach to the one we describe below can also be used to estimate the entire joint density (3.6). For simplicity, we assume that in (3.6), there exist functions ψ and ζ such that $\psi = \psi_{j,k}$ for all (j,k) and $\zeta = \zeta_{j,c}$ for all (j,c), and that $\mu_j = 0$. For $x = (x_1, \ldots, x_p)^\top$ and $w = (w_1, \ldots, w_q)^\top$ the conditional density can thus be expressed as

$$f_j^{g,*}(x_j | x_1, \ldots, x_p, w) \propto \exp\left(\sum_{j=1}^{p} \left\langle \alpha_{j,k}^{g,*}, \phi(w) \right\rangle \psi(x_j, x_k) + \sum_{c=1}^{d} \theta_{j,c}^{g,*} \zeta\left(x_j, \phi_c(w)\right) \right), \quad (3.17)$$

where the proportionality is up to a normalizing constant that does not depend on x_j.

We first explicitly define the score matching loss for the conditional density function (3.17). Let $\boldsymbol{\alpha}_j^{g,*} = \left((\alpha_{j,1}^{g,*})^\top, \ldots, (\alpha_{j,p}^{g,*})^\top\right)^\top$, and similarly let $\boldsymbol{\theta}_j^{g,*} = (\theta_{j,1}^{g,*}, \ldots, \theta_{j,p}^{g,*})^\top$. Let $\dot{\psi}$ and $\ddot{\psi}$ denote the first and second derivatives of ψ with respect to x_j, and similarly, let $\dot{\zeta}$ and $\ddot{\zeta}$ denote the first and second derivatives of ζ with respect to x_j. We define a non-negative function $v_j : \mathbb{R}_+ \to \mathbb{R}_+$, and let $\dot{v}_j$ denote the first derivative of v_j. Then for candidate parameters $\boldsymbol{\alpha}_j = \left(\alpha_{j,1}^\top, \ldots, \alpha_{j,p}^\top\right)^\top$ and $\boldsymbol{\theta}_j = (\theta_{j,1}, \ldots, \theta_{j,d})^\top$, the empirical generalized score matching loss for the conditional density of X_j^g given all other nodes and the covariates can be expressed as

$$\begin{aligned} L_{n,j}^g(\boldsymbol{\alpha}, \boldsymbol{\theta}) = {} & \frac{1}{2n^g} \sum_{i=1}^{n^g} v_j\left(X_{i,j}^g\right) \left\{ \sum_{k=1}^{p} \left\langle \alpha_{j,k}, \phi\left(W_i^g\right) \right\rangle \dot{\psi}\left(X_{i,j}^g, X_{i,k}^g\right) + \sum_{c=1}^{d} \theta_{j,c} \dot{\zeta}\left(X_{i,j}^g, \phi_c\left(W_i^g\right)\right) \right\}^2 + \\ & \frac{1}{n^g} \sum_{i=1}^{n^g} v_j\left(X_{i,j}^g\right) \left\{ \sum_{k=1}^{p} \left\langle \alpha_{j,k}, \phi\left(W_i^g\right) \right\rangle \ddot{\psi}\left(X_{i,j}^g, X_{i,k}^g\right) + \sum_{c=1}^{d} \theta_{j,c} \ddot{\zeta}\left(X_{i,j}^g, \phi_c\left(W_i^g\right)\right) \right\} + \\ & \frac{1}{n^g} \sum_{i=1}^{n^g} \dot{v}_j\left(X_{i,j}^g\right) \left\{ \sum_{k=1}^{p} \left\langle \alpha_{j,k}, \phi\left(W_i^g\right) \right\rangle \dot{\psi}\left(X_{i,j}^g, X_{i,k}^g\right) + \sum_{c=1}^{d} \theta_{j,c} \dot{\zeta}\left(X_{i,j}^g, \phi_c\left(W_i^g\right)\right) \right\}. \end{aligned} \quad (3.18)$$

The true parameters $\boldsymbol{\alpha}_j^{g,*}$ and $\boldsymbol{\theta}_j^{g,*}$ can characterized as the minimizers of the population score matching loss $E\left[L_{n,j}^g\left(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j\right)\right]$, as discussed in Section 3.4.1.

The loss function in (3.18) is quadratic in parameters $\boldsymbol{\alpha}_j^{g,*}$ and $\boldsymbol{\theta}_j^{g,*}$ and can thus be solved efficiently. When the sample size n^g is much larger than the number of unknown parameters $(p+1)d$, one can estimate $\boldsymbol{\alpha}_j^{g,*}$ and $\boldsymbol{\theta}_j^{g,*}$ by simply minimizing $L_{n,j}^g$ with respect to the unknown parameters. The empirical loss function is quadratic in $(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j)$, so the minimizer of the loss is available in closed form and can be computed efficiently. Moreover, we can readily establish asymptotic normality of the parameter estimates using results from classical M-estimation theory van der Vaart (2000, Chapter 5). To avoid including cumbersome notation, we reserve the details for Appendix B.2.2.

When the sample size is smaller than the number of parameters, the minimizer of $L_{n,j}^g$ is no longer well-defined. Similar to Section 3.3.2, we use regularization to obtain a consistent estimator in the high-dimensional setting. We define the ℓ_2-regularized generalized score matching estimator as

$$\left(\tilde{\boldsymbol{\alpha}}_j^g, \tilde{\boldsymbol{\theta}}_j^g\right) = \underset{\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j}{\operatorname{argmin}} \; L_{n,j}^g(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j) + \lambda \sum_{j=1}^{p} \|\alpha_{j,k}\|_2 \,, \tag{3.19}$$

where $\lambda > 0$ is a tuning parameter. Similar to the group LASSO estimator (3.12), the regularization term in (3.19) induces sparsity in the estimate $\tilde{\boldsymbol{\alpha}}_j^g$ and sets some $\tilde{\alpha}_{j,k}^g$ to be exactly zero. The tuning parameter controls the level of sparsity, where more vectors $\tilde{\alpha}_{j,k}^g$ are zero for higher λ. In Appendix B.2.3, we establish consistency of the regularized score matching estimator assuming sparsity of $\tilde{\boldsymbol{\alpha}}_j^g$ and some additional regularity conditions. The convergence rate and conditions on the tuning parameter λ depend on the probability distribution. We show that for some distributions, the rate $\sum_{j=1}^p \|\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\|_2 = O_P(\{n \log(p)\}^{1/2})$ can be achieved with $\lambda \propto \{\log(p)/n\}^{1/2}$.

As is the case for the group LASSO estimator, the regularized score matching estimator has an intractable limiting distribution because its bias and standard error diminish at the same rate. We can obtain an asymptotically normal estimate by subtracting from the initial estimate an estimate of the bias. In Appendix B.2.4, we construct such a bias-corrected

estimate $\check{\alpha}^g_{j,k}$ that, for sufficiently large n^g, satisfies

$$\check{\alpha}^g_{j,k} \sim N\left(\alpha^{g,*}_{j,k}, \Omega^g_{j,k}\right),$$

for a positive definite matrix $\Omega^g_{j,k}$. Given bias-corrected estimates and a consistent estimate $\check{\Omega}^g_{j,k}$ of $\Omega^g_{j,k}$, we can test the null hypothesis (3.5) using the test statistic

$$S_{j,k} = \left(\check{\alpha}^{\mathrm{I}}_{j,k} - \check{\alpha}^{\mathrm{II}}_{j,k}\right)^\top \left(\check{\Omega}^{\mathrm{I}}_{j,k} + \check{\Omega}^{\mathrm{II}}_{j,k}\right)^{-1} \left(\check{\alpha}^{\mathrm{I}}_{j,k} - \check{\alpha}^{\mathrm{II}}_{j,k}\right).$$

Under the null hypothesis, the test statistic follows a chi-squared distribution with d degrees of freedom.

3.5 Numerical studies

In this section, we examine the performance of our proposed test in a simulation study. We consider the neighborhood selection approach described in Section 3.3. Our simulation study has three objectives: (1) to assess the stability of our estimators for the covariate-dependent networks, (2) to examine the effect of sample size on statistical power and type-1 error control, and (3) to illustrate that failing to adjust for covariates can in some settings result in poor type-1 error control or reduced statistical power.

3.5.1 Implementation

We first discuss implementation of the neighborhood selection approach. The group LASSO estimate (3.12) does not exist in closed form, in contrast to the ordinary least squares estimate (3.11). To solve (3.12), we use the efficient algorithm implemented in the publicly available R package `gglasso` (Yang and Zou, 2015).

The group LASSO estimator requires selection of a tuning parameter λ, which controls the sparsity of the estimate. We select the tuning parameter by performing K-fold cross-validation, using $K = 10$ folds. While cross-validation is the gold standard for tuning parameter selection and typically leads to good empirical performance, we note that the group LASSO estimator estimator may converge slightly slower than the oracle rate when λ

is selected via cross-validation (Homrighausen and McDonald, 2017). Since the selection of λ is sensitive to the scale of the columns of $\mathcal{V}_k^g$ in (3.10), we scale the columns by their standard deviations prior to cross-validating. After fitting the group LASSO with the selected tuning parameter, we convert the estimates back to their original scale by dividing the estimates by the standard deviations of the columns of $\mathcal{V}_k^g$.

3.5.2 Simulation setting

In what follows, we describe our simulation setting. In short, we generate data from the varying coefficient model (3.8), where we treat nodes 1 through $(p-1)$ as predictors, and treat node p as the response. We first randomly generate data for nodes 1 through $(p-1)$ in groups I and II from the same multivariate normal distribution. We then construct $\eta_{j,k}^{g,*}$ and generate data for two covariates $W_i^g = (W_{i,1}^g, W_{i,2}^g)^\top$ so that one covariate acts as a confounding variable, and the other covariate should improve statistical power to detect differential associations after adjustment.

To simulate data for nodes 1 through $(p-1)$, we first generate a random graph with $(p-1)$ nodes and an edge density of .05 from a power law distribution with power parameter 5 (Newman, 2003). Denoting the edge set of the graph by E, we generate the $(p-1)\times(p-1)$ matrix Θ as

$$\Theta_{j,k} = \begin{cases} 0 & (j,k) \notin E \\ .5 & (j,k) \in E \text{ with } 50\% \text{ probability} \\ -.5 & (j,k) \in E \text{ with } 50\% \text{ probability} \end{cases},$$

with $\Theta_{j,k} = \Theta_{k,j}$. Defining by a^* the smallest eigenvalue of Θ, we set $\Sigma = (\Theta - (a^* - .1)I)^{-1}$, where I is the identity matrix. We then draw $(X_{i,1}^g, \ldots, X_{i,p-1}^g)^\top$ from a multivariate normal distribution with mean zero and covariance Σ for $i = 1, \ldots, n^g$ for each group g.

We generate $W_{i,1}^{\mathrm{I}}$ from a Beta$(3/2, 1)$ distribution and $W_{i,1}^{\mathrm{II}}$ from a Beta$(1, 3/2)$ distribution. We center and scale both $W_{i,1}^{\mathrm{I}}$ and $W_{i,1}^{\mathrm{II}}$ to the $(-1, 1)$ interval. We generate $W_{i,2}^{\mathrm{I}}$ and $W_{i,2}^{\mathrm{II}}$ each from a Uniform$(-1, 1)$ distribution.

We consider two different choices for the varying coefficient functions $\eta_{j,k}^{g,*}$:

- *Linear Polynomial:*

$$\begin{aligned}
\eta_{p,1}^{\mathrm{I},*}(w_1, w_2) &= .5 + .5w_1; & \eta_{p,1}^{\mathrm{II},*}(w_1, w_2) &= .5 + .5w_1 \\
\eta_{p,2}^{\mathrm{I},*}(w_1, w_2) &= .5 + .25w_2; & \eta_{p,2}^{\mathrm{II},*}(w_1, w_2) &= .5 + .75w_2 \\
\eta_{p,3}^{\mathrm{I},*}(w_1, w_2) &= 0; & \eta_{p,3}^{\mathrm{II},*}(w_1, w_2) &= .5,
\end{aligned}$$

  and $\eta_{p,k}^{g,*} = 0$ for $k \geq 4$.

- *Cubic Polynomial:*

$$\begin{aligned}
\eta_{p,1}^{\mathrm{I},*}(w_1, w_2) &= .5 + .5\left(w_1 + w_1^2 + w_1^3\right); & \eta_{p,1}^{\mathrm{II},*}(w_1, w_2) &= .5 + .5\left(w_1 + w_1^2 + w_1^3\right) \\
\eta_{p,2}^{\mathrm{I},*}(w_1, w_2) &= .5 + .25\left(w_2 + w_2^3\right); & \eta_{p,2}^{\mathrm{II},*}(w_1, w_2) &= .5 + .75\left(w_2 + w_2^3\right) \\
\eta_{p,3}^{\mathrm{I},*}(w_1, w_2) &= 0; & \eta_{p,3}^{\mathrm{II},*}(w_1, w_2) &= .5,
\end{aligned}$$

  and $\eta_{p,k}^{g,*} = 0$ for $k \geq 4$.

The first covariate $W_{i,1}^g$ confounds the association between nodes p and 1. The distribution of $W_{i,1}^g$ depends on group membership, and $W_{i,1}^g$ affects the association between genes p and 1. However, $\eta_{p,1}^{\mathrm{I},*}(w) = \eta_{p,1}^{\mathrm{II},*}(w)$ for all w. Thus, $G_{p,1}^0$ in (3.5) holds while $H_{p,1}^0$ in (3.1) fails, as depicted in Figure 3.1a. Failing to adjust for W_1^g should therefore result in an inflated type-1 error rate for the hypothesis $G_{p,1}^0$. Adjusting for the second covariate $W_{i,2}^g$ should improve the power to detect the differential connection between nodes p and 2. We have constructed $\eta_{p,2}^{g,*}$ so that $E\left[\eta^{\mathrm{I},*}\left(W^{\mathrm{I}}\right)\right] = E\left[\eta^{\mathrm{II},*}\left(W^{\mathrm{II}}\right)\right]$, though the association between nodes p and 2 depends more strongly on W^g in group II than in group I. Thus, $H_{p,2}^0$ holds while $G_{p,2}^0$ fails, as depicted in Figure 3.1b. The association between nodes p and 3 does not depend on either covariate, though the association differs by group. Thus, one should be able to identify a differential connection using either the adjusted or unadjusted test. Node p is conditionally independent of all other nodes in both groups.

For $i = 1, \ldots, n^g$, we generate $X^g_{i,p}$ as

$$X^g_{i,p} = \sum_{k \neq p} \eta^g_{j,k}\left(W^g_i\right) X^g_{i,k} + \epsilon^g_i,$$

where ϵ^g_i follows a normal distribution with zero mean and unit variance. We use balanced sample sizes $n^{\mathrm{I}} = n^{\mathrm{II}} = n$ and consider $n \in \{80, 160, 240\}$. We set the number of nodes $p = 40$. The graph for nodes 1 through $(p-1)$ contains 15 edges. Leaving Σ fixed, we generate 400 random data sets following the above approach.

We consider two choices of the basis expansion ϕ:

1. Linear basis: $\phi(w_1, w_2) = \begin{pmatrix} 1 & w_1 & w_2 \end{pmatrix}^\top$;
2. Cubic polynomial basis: $\phi(w_1, w_2) = \begin{pmatrix} 1 & w_1 & w_1^2 & w_1^3 & w_2 & w_2^2 & w_2^3 \end{pmatrix}^\top$.

Using a linear basis, $d = 3$, and model (3.8) has 117 parameters. With the cubic polynomial basis, $d = 7$, and there are 273 parameters.

We compare our proposed methodology with the approach for differential network analysis without covariate adjustment described in Section 3.3.1. In the unadjusted analysis, ordinary least squares estimation is justified because although $(p-1)d$ is large with respect to n, $(p-1)$ is smaller than n.

3.5.3 Simulation results

Figure 3.2 shows the Monte Carlo estimates of the expected ℓ_2 error for the de-biased group LASSO estimates $\check{\alpha}^g_{p,k}$, $E\left[\left\|d^{-1}\left(\check{\alpha}^g_{p,k} - \alpha^{g,*}_{p,k}\right)\right\|_2\right]$, for $k = 1, \ldots, (p-1)$. We only report the ℓ_2 error when the basis ϕ is correctly specified for the varying coefficient function $\eta^{g,*}_{p,k}$ — that is, when ϕ is linear basis, and $\eta^{g,*}_{p,k}$ is a linear function or when ϕ is a cubic basis, and $\eta^{g,*}_{p,k}$ is a cubic function. In both the linear and cubic polynomial settings, the average ℓ_2 estimation error for $\alpha^{g,*}_{p,k}$ decreases with the sample size for all k, as expected. We also find that in small samples, the estimation error is substantially lower in the linear setting than in the cubic setting. This suggests that estimates are less stable in more complex models.

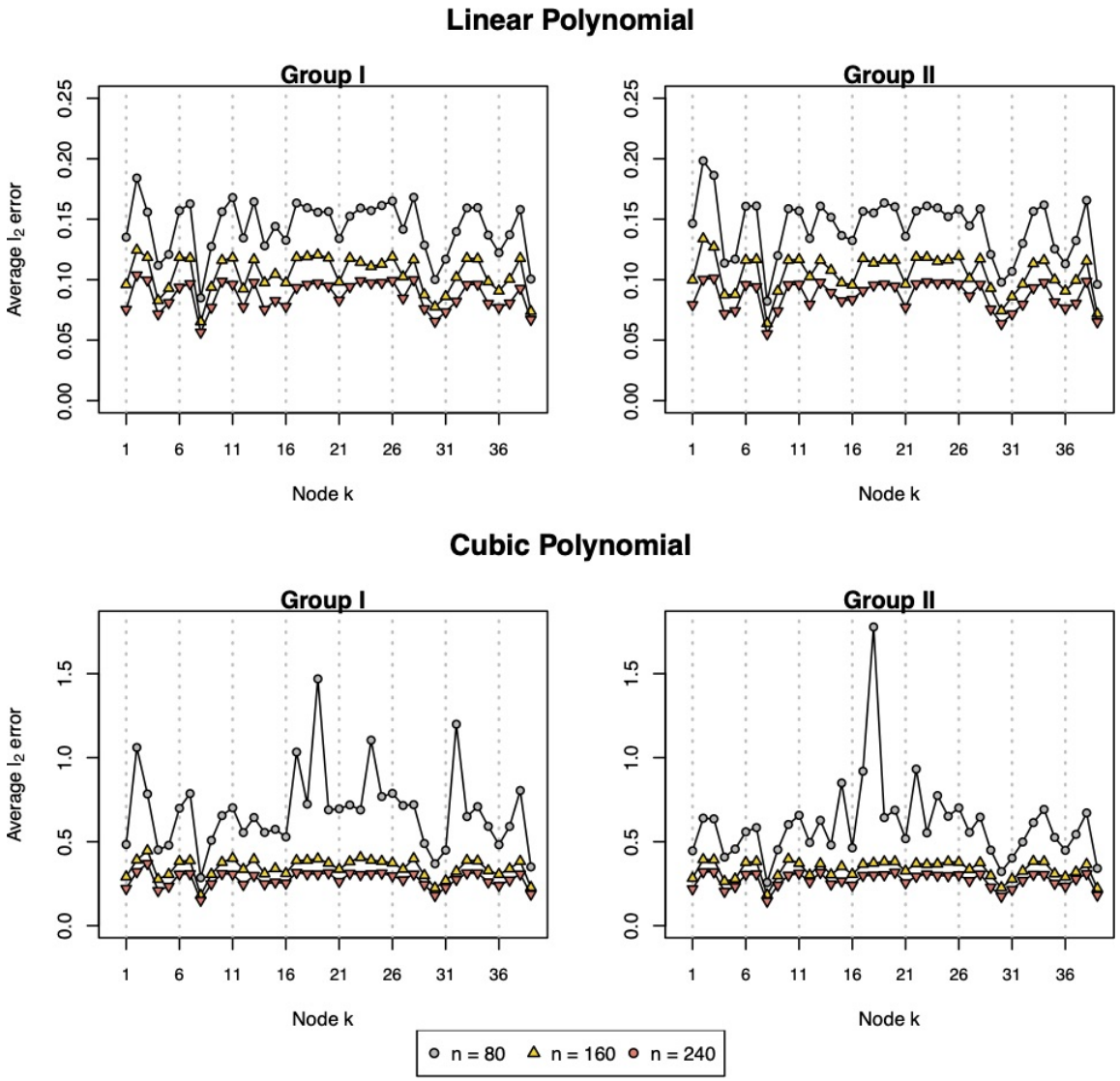


Figure 3.2: Monte Carlo estimates of expected ℓ_2 error, $\mathbb{E}\left[\left\|d^{-1}\left(\breve{\alpha}^g_{p,k} - \alpha^{g,*}_{p,k}\right)\right\|_2\right]$, for $k = 1, \ldots, 39$. The linear polynomial plots display the ℓ_2 error when $\eta^{g,*}_{j,k}$ is a linear function, and ϕ is a linear basis. The cubic polynomial plots display the ℓ_2 error when $\eta^{g,*}_{j,k}$ is a cubic polynomial, and ϕ is a cubic basis.

In Table 3.1, we report Monte Carlo estimates of the probability of rejecting $G^0_{p,k}$, the null hypothesis that nodes p and k are not differentially connected given W^g, for $k = 1$,

$k = 2$, $k = 3$, and $k \geq 4$, using both the adjusted and unadjusted tests at the significance level $\mathsf{a} = .05$. As the purpose of the simulation study is to examine the behavior of the edge-wise test, we do not perform a multiple testing correction.

For $k = 1$ (i.e., when $H^0_{p,k}$ fails, but $G^0_{p,k}$ holds), the unadjusted test is anti-conservative, and the probability of falsely rejecting $G^0_{p,k}$ increases with the sample size. When an adjusted test is performed using a linear basis, and when $\eta^{g,*}_{p,1}$ is linear, the type-1 error rate is slightly inflated but appears to approach the nominal level of .05 as the sample size increases. However, when $\eta^{g,*}_{p,1}$ is a cubic function, and the linear basis is mis-specified, the type-1 error rate is inflated, though it is still slightly lower than that of unadjusted test. For both specifications of $\eta^{g,*}_{p,1}$, the covariate-adjusted test controls the type-1 error rate near the nominal level when a cubic polynomial basis is used. For $k = 2$, (i.e., when $H^0_{p,k}$ holds, but $G^0_{p,k}$ fails), the unadjusted test exhibits low power to detect differential associations. The adjusted test provides greatly improved power when either a linear or cubic basis is used. For $k = 3$, (i.e., when both $H^0_{p,k}$ and $G^0_{p,k}$ fail), the unadjusted test and both adjusted tests are well-powered against the null. For $k \geq 4$ (i.e., when genes p and k are conditionally independent in both groups), the unadjusted test and the adjusted test with a linear basis both control the type-1 error near the nominal level. However, the covariate-adjusted test is conservative when a cubic basis is used.

The simulation results corroborate our expectations and suggest that there are potential benefits to covariate adjustment. We find that when the sample size is large, the covariate-adjusted test behaves reasonably well with either choice of basis function. However, in small samples, the covariate-adjusted test is somewhat imprecise, and the type-1 error rate can be slightly above or below the nominal level. Practitioners should therefore exercise caution when using our proposed methodology in small samples.

3.6 Data example

Breast cancer classification based on expression of estrogen receptor hormone (ER) is prognostic of clinical outcomes. Breast cancers can be classified as estrogen receptor positive

		Unadjusted			Linear Adjustment			Cubic Adjustment		
		$n = 80$	$n = 160$	$n = 240$	$n = 80$	$n = 160$	$n = 240$	$n = 80$	$n = 160$	$n = 240$
Linear $\eta^{g,*}_{p,k}$	$k = 1$	0.15	0.278	0.385	0.13	0.09	0.072	0.04	0.062	0.05
	$k = 2$	0.042	0.078	0.045	0.27	0.532	0.73	0.08	0.27	0.52
	$k = 3$	0.48	0.912	0.988	0.605	0.922	0.965	0.218	0.738	0.902
	$k \geq 4$	0.052	0.054	0.053	0.045	0.048	0.048	0.009	0.017	0.025
Cubic $\eta^{g,*}_{p,k}$	$k = 1$	0.21	0.505	0.668	0.358	0.315	0.342	0.07	0.055	0.07
	$k = 2$	0.052	0.068	0.082	0.6	0.882	0.975	0.195	0.73	0.93
	$k = 3$	0.408	0.84	0.978	0.55	0.898	0.982	0.202	0.772	0.945
	$k \geq 4$	0.056	0.05	0.054	0.053	0.054	0.052	0.009	0.02	0.027

Table 3.1: Monte Carlo estimates of probability of rejecting $G^0_{p,k}$, the null hypothesis that nodes p and k are not differentially connected, given W^g. All tests are performed at the significance level $\mathsf{a} = .05$, and no multiple testing correction is performed.

(ER+) and estrogen receptor negative (ER-), with approximately 70% of breast cancers being ER+ (Lumachi et al., 2013). In ER+ breast cancer, the cancer cells require estrogen to grow; this has been shown to be associated with positive clinical outcomes, compared with ER- breast cancer (Carey et al., 2006). Identifying differences between the biological pathways of ER+ and ER- breast cancers can be helpful for understanding the underlying disease mechanisms.

It is has been shown that age is associated with ER status and that age can be associated with gene expression (Khan et al., 1998; Yang et al., 2015). This warrants consideration of age as an adjustment variable in a comparison of gene co-expression networks between ER groups, as.

We perform an age-adjusted differential analysis of the ER+ and ER- breast cancer networks, using publicly available data from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). We obtain clinical measurements and gene expression data from a total of 806 ER+ patients and 237 ER- patients. We consider the set of $p = 145$ genes in the Kyoto

Encyclopedia of Genes and Genomes (KEGG) breast cancer pathway (Kanehisa and Goto, 2000), and adjust for age as our only covariate. The average age in the ER+ plus group is 59.3 years (SD = 13.3), and the average age in the ER- group is 55.9 years (SD = 12.4). We use a linear basis for covariate adjustment. In the ER+ group, the sample size is considerably larger than the number of the parameters, so we can fit the varying coefficient model (3.8) using ordinary least squares. We use the de-biased group LASSO to estimate the network for the ER- group because the sample size is smaller than the number of model parameters. We compare the results from the covariate-adjusted analysis with the unadjusted approach described in Section 3.3.1.

To assess for differential connectivity between any two nodes j and k, we can either treat node j or node k as the response in the varying coefficient model (3.8). We can then test either of the hypotheses $G^0_{j,k} : \alpha^{\mathrm{I},*}_{j,k} = \alpha^{\mathrm{II},*}_{j,k}$ or $G^0_{k,j} : \alpha^{\mathrm{I},*}_{k,j} = \alpha^{\mathrm{II},*}_{k,j}$. We use a Bonferroni adjustment and set our p-value for the test for differential connectivity between nodes j and k as two times the minimum of the p-values for the tests of $G^0_{j,k}$ and $G^0_{k,j}$. While we use this conservative approach due to its simplicity, we could alternatively construct a simultaneous test of $G^0_{j,k}$ and $G^0_{k,j}$ by characterizing the joint limiting distribution of $\left((\check{\alpha}^g_{j,k})^\top, (\check{\alpha}^g_{k,j})^\top\right)^\top$.

Our objective is to identify all pairs of differentially connected genes, so we need to adjust for the fact that we perform a separate hypothesis test for each gene pair. We account for multiplicity by controlling the false discovery rate at the level $\mathsf{a} = .05$ using the Benjamini-Yekutieli method (Benjamini and Yekutieli, 2001).

The differential networks obtained from the unadjusted and adjusted analyses are substantially different. We report 79 differentially connected edges from the adjusted analysis (shown in Figure 3.3), compared to two edges from the unadjusted analysis. This suggests it is possible that relationship between the gene co-expression network and age differs by ER group.

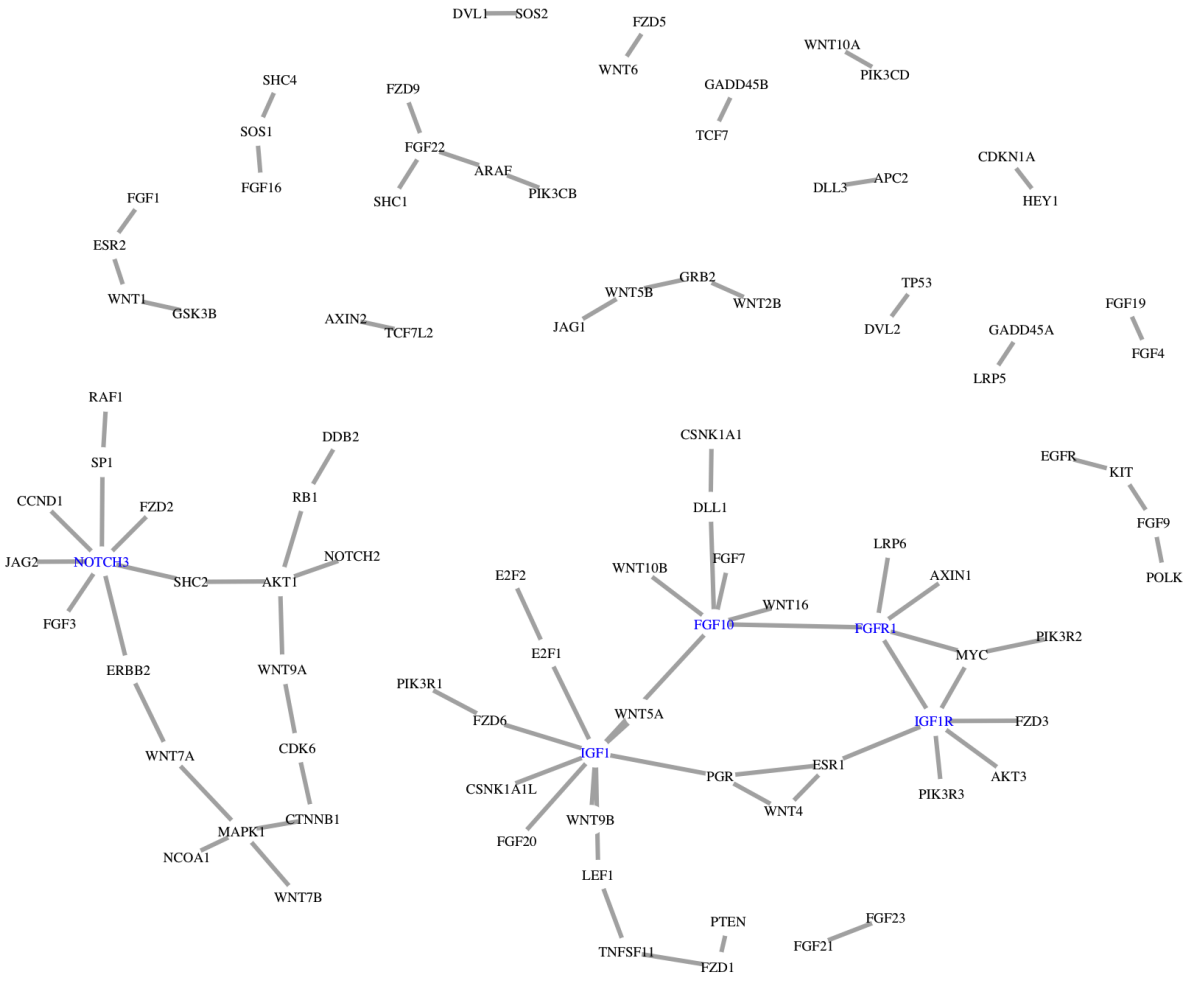


Figure 3.3: Differential breast cancer network by estrogen receptor status from covariate-adjusted analysis. Nodes with at least five differentially connected neighbors are colored blue. The false discovery rate is controlled at .05.

3.7 Discussion

In this chapter, we have addressed challenges that arise when performing differential network analysis (Shojaie, 2020) in the setting where the network depends on covariates. Using both

synthetic and real data, we showed that accounting for covariates can result in better control of type-1 error and improved power.

We propose a parsimonious approach for covariate adjustment in differential network analysis. A number of improvements and extensions can be made to our current work. First, while this chapter focuses on differential network analysis in exponential family models, our framework can be applied to other models where conditional dependence between any pair of nodes can be represented by a single scalar parameter. This includes semi-parametric models such as the nonparanormal model (Liu et al., 2009), as well as distributions defined over complex domains, which can be modeled using the generalized score matching framework (Yu et al., 2021). Additionally, we only discuss testing edge-wise differences between the networks, though testing differences between sub-networks may also be of interest. When the sub-networks are low-dimensional, one can construct a chi-squared test using similar test statistics as presented in Section 3.3 and Section 3.4 because joint asymptotic normality of a low-dimensional set of the estimators $\check{\alpha}^g_{j,k}$ can be readily established. Such an approach is not applicable to high-dimensional sub-networks, but it may be possible to construct a calibrated test using recent results on simultaneous inference in high-dimensional models (Zhang and Cheng, 2017; Yu et al., 2020). We can also improve the statistical efficiency of the network estimates by considering joint estimation procedures that borrow information across groups (Guo et al., 2011; Danaher et al., 2014; Saegusa and Shojaie, 2016). Finally, we assume that the relationship between the network and the covariates can be represented by a low-dimensional basis expansion. Investigating nonparametric approaches that relax this assumption can be a fruitful area of research.

Chapter 4

INFERENCE ON FUNCTION-VALUED PARAMETERS USING A RESTRICTED SCORE TEST

4.1 Introduction

It is often the case that the estimand of scientific interest in a given application is an unknown function, either in its entirety or through its evaluation at one or several points in its domain. As a parameter of the underlying data-generating mechanism, such function is either global — such as a distribution or quantile function — or local — such as a density, conditional mean or hazard function. When sufficiently strong parametric assumptions are made, inference on such an estimand, be it local or global, is usually straightforward, and parametric rates of estimation are achievable. For instance, it is common to assume that a regression function is linear or polynomial of a fixed degree, and inference then only involves a finite set of unknown regression coefficients, which can be readily estimated at the parametric rate under weak conditions. However, such restrictive modeling assumptions bring the risk of invalid inference due to model misspecification. This fact has motivated investigators to instead rely on nonparametric or semiparametric models, for which the risk of model misspecification is reduced.

In nonparametric and semiparametric models, whether the parameter is local or global determines whether regular parametric-rate estimators exist for the unknown function, and therefore, how challenging calibrated inference is to achieve. When the unknown function is a global parameter of the data-generating mechanism, parametric-rate inference is possible, and there is a well-established efficiency theory that characterizes the large-sample behavior of optimal estimators (see, e.g., the text by (Bickel et al., 1998) for a comprehensive exposition). There also exist several constructive approaches for obtaining such estimators,

including one-step debiasing procedures (Pfanzagl, 1982), estimating equations (van der Laan and Robins, 2003; Chernozhukov et al., 2018), and targeted minimum loss-based estimation (van der Laan and Rose, 2011).

In contrast, when the unknown function is a local parameter of the data-generating mechanism, there usually does not exist any regular parametric-rate estimator. While estimation strategies abound for this setting, there are few formal approaches for inference (e.g., construction of confidence sets and hypothesis tests) based on these strategies since studying the limiting distribution of such estimators is usually challenging. For example, it is often the case that the bias of an estimator obtained by minimizing an empirical risk criterion tends to zero at the same rate as its standard error. To address this bias and facilitate inference, several approaches have been proposed, many in the context of kernel smoothing. One common approach consists of constructing a data-driven bias correction (see, e.g., Hardle and Marron, 1991; Eubank and Speckman, 1993; Sun et al., 1994; Calonico et al., 2018; Lu et al., 2020). Another approach consists of undersmoothing, that is, selecting tuning parameter values that deliberately inflate the variance in order to deflate the bias of the estimator, even though such choice results in a suboptimal risk value (see, e.g., Hall, 1991, 1992; Neumann et al., 1995). Both approaches typically require a characterization of the bias of the considered estimator, which often has a complex form and is difficult to estimate. They can also be sensitive to tuning parameter selection and difficult to implement in practice. As an alternative, Hall et al. (2013) suggests a bootstrap-based algorithm for pointwise inference, whereas van der Laan et al. (2018) proposes to use targeted minimum loss-based estimation on a sequence of decreasingly-regularized modifications of the original estimand. However, neither approach appears to be directly applicable when uniform coverage or simultaneous testing is of interest.

In many settings, the estimand of interest can be represented as the minimizer of a population risk functional. Here, we use this representation to develop a novel general framework for inference for function-valued parameters in nonparametric and semiparametric models. We propose a test based on assessing the feasibility of the null parameter value by evaluating

how greatly the derivative of the risk functional at this value differs from zero. Our test is an infinite-dimensional extension of the classical Rao score test Rao (1948), and in fact, reduces to this test when the model considered is finite-dimensional. By inverting the proposed test, we obtain uniform confidence bands for the unknown function of interest or its evaluation on a set. In contrast to existing approaches applicable in infinite-dimensional models, our proposal does not require estimation of the unknown function itself, thereby circumventing the difficulties usually caused by the bias of existing estimators. The framework we propose is quite general, and applies equally to classical parameters, such as density and conditional mean functions and to more complicated parameters, so long as the parameter of interest is a population risk minimizer. Our proposal is also flexible, requiring few assumptions about the data-generating mechanism.

The rest of the chapter is organized as follows. In Section 4.2, we present a high-level sketch of our proposed framework for inference. In Section 4.3, we discuss estimation of risk functional derivatives, which play a critical role in our proposed approach. We present the theoretical results supporting our approach in Section 4.4, and discuss practical considerations arising in its implementation in Section 4.5. We evaluate the operating characteristics of the proposed method in Section 4.6, and apply it to data from the 1987 National Medical Expenditures Survey in Section 4.7. We provide concluding remarks in Section 4.8.

4.2 Overview of the proposed framework

4.2.1 Preliminaries

We begin by introducing some definitions and the notation used throughout. Let $Z_1, Z_2, \ldots, Z_n$ represent independent random vectors drawn from a distribution P_0 known only to reside in a potentially rich statistical model $\mathcal{M}$, and denote by $\mathcal{Z}$ the sample space corresponding to P_0. Suppose that Θ is a given function class and $P \mapsto \theta_P \in \Theta$ is a function-valued parameter mapping defined over $\mathcal{M}$. We are interested in making inference on $\theta_0 := \theta_{P_0}$. Suppose that for each $P \in \mathcal{M}$ there exists a P-risk functional $R_P : \Theta \to \mathbb{R}$ such that

$\theta_P = \text{argmin}_{\theta \in \Theta} R_P(\theta)$. In particular, defining the shorthand notation $R_0 := R_{P_0}$, this allows us to write the representation $\theta_0 = \text{argmin}_{\theta \in \Theta} R_0(\theta)$. In many cases, the risk functional has the simpler form

$$R_P(\theta) = \mathbb{E}_P[\ell_P(\theta, Z)] \tag{4.1}$$

for some loss function $\ell_P : \Theta \times \mathcal{Z} \to \mathbb{R}$ indexed by P, though this is never required below.

4.2.2 Working examples

Minimizers of risk functionals arise naturally in many problems. Before describing our proposed approach, we describe the two working examples we will refer to throughout the manuscript. Both examples pertain to conditional mean functions. Let $Z := (X, W, Y)$, where $Y \in \mathbb{R}$ is the response variable and $(X, W) \in \mathbb{R} \times \mathbb{R}^d$ represents a covariate vector. In the first example, we will consider inference on the conditional mean function $\theta_0 : x \mapsto \mathbb{E}_0(Y \mid X = x)$ under a nonparametric model, where here and below $\mathbb{E}_0$ denotes expectation under P_0. Because we can express θ_0 as the minimizer of the least-squares risk, that is,

$$\theta_0 = \underset{\theta \in \Theta}{\text{argmin}}\, \mathbb{E}_0\left[\{Y - \theta(X)\}^2\right] \tag{4.2}$$

with Θ taken to be the space $L_2(P_0)$ of P_0-square-integrable real-valued functions defined on the support of X, this corresponds to using the P-risk functional $R_P(\theta) := \mathbb{E}_P\left[\{Y - \theta(X)\}^2\right]$. We readily see that $P \mapsto R_P(\theta)$ is a linear functional. In fact, this risk functional has the form (4.1) with $\ell_P : (\theta, z) \mapsto \{y - \theta(x)\}^2$, where ℓ_P does not depend on P at all.

In the second example, we will consider inference on θ_0 under a partially additive mean model that enforces the structure $\mathbb{E}_0(Y \mid X = x, W = w) = \theta_0(x) + f_0(w)$ for unknown functions θ_0 and f_0 with condition $\mathbb{E}_0[\theta_0(X)] = 0$ imposed for identifiability. The partially linear model is a special case of the partially additive model under which θ_0 must also be linear (Robinson, 1988). The parameter value θ_0 facilitates a quantification of the association

between an exposure X and outcome Y after adjustment for a vector W of potential confounders: specifically, $\theta_0(x_1) - \theta_0(x_0)$ represents the difference in mean outcome between two subpopulations of individuals with exposure levels x_1 and x_0 but same level of confounding factors. Similarly as before, θ_0 can be expressed as a minimizer of a population least-squares risk, that is,

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathbb{E}_0 \left[\{ Y - f_0(W) - \theta(X) \}^2 \right] \tag{4.3}$$

with Θ now taken to be the subset $L_2^0(P_0)$ of elements of $L_2(P_0)$ with P_0-mean zero. This corresponds to using the P-risk functional $R_P(\theta) := \mathbb{E}_P \left[\{ Y - f_P(W) - \theta(X) \}^2 \right]$, where the nuisance function f_P is such that $\mathbb{E}_P \left(Y \mid X = x, W = w \right) = \theta_P(x) + f_P(w)$ for each x and w. In this case, the risk functional also has the form (4.1) with $\ell_P : (\theta, z) \mapsto \{ y - f_P(w) - \theta(x) \}^2$, a loss function that depends on P through f_P. As such, in this case, $\theta \mapsto R_P(\theta)$ is not a linear functional.

4.2.3 General inferential strategy

Our objective is to conduct formal inference for an arbitrary population risk minimizer as defined above. We begin by constructing a level $\alpha \in (0, 1)$ test of the hypothesis

$$H_0 : \theta_0 = \theta_* \ ,$$

where $\theta_* \in \Theta$ is a pre-specified null parameter value, against the complement hypothesis $H_1 : \theta_0 \neq \theta_*$. Then, by inverting this test, we derive a confidence region $\mathcal{C}_n$ for θ_0, that is, we obtain a random set $\mathcal{C}_n = \mathcal{C}_n(Z_1, Z_2, \ldots, Z_n) \subset \Theta$ that contains the population risk minimizer with probability at least $1 - \alpha$ as sample size n tends to infinity:

$$\liminf_{n \to \infty} \mathbb{P}_0 \left(\theta_0 \in \mathcal{C}_n \right) \geq 1 - \alpha \ .$$

Our proposal is closely related to the classical (parametric) score test of Rao (1948), which we briefly review. Suppose that Θ is the collection of linear functions $\left\{ \theta_\beta : z \mapsto z^\top \beta : \beta \in \mathbb{R}^p \right\}$, a set indexed by the vector β of finite dimension p. In the finite-dimensional setting, for

any null value $\beta_* \in \mathbb{R}^p$ of the index parameter, the classical score test assesses whether the null function θ_{β_*} is a population risk minimizer by determining whether there is empirical evidence to suggest that the derivative of the risk function $\beta \mapsto R_0(\theta_\beta)$ evaluated at β_* is zero. If the derivative is nonzero, θ_{β_*} cannot be a population risk minimizer. This approach is based on studying local perturbations of R_0 in a neighborhood of the null value θ_{β_*} along a finite-dimensional collection of directions. As such, it does not rely upon estimation of the population risk minimizer θ_0. We find this approach appealing because it allows one to conduct inference even when it is difficult to construct an estimator of θ_0 with a tractable limiting distribution.

Our proposal generalizes the classical score test to the infinite-dimensional setting. For this generalization, we will require a proper notion of differentiability of R_0 over the infinite-dimensional space Θ. For simplicity, suppose that Θ is a convex space. We will say that the population risk functional $R_0 : \Theta \to \mathbb{R}$ is Gâteaux differentiable at $\theta = \theta_*$ provided that, for each direction $h \in \mathcal{H}(\theta_*) := \{\theta - \theta_* : \theta \in \Theta\}$, the Gâteaux derivative

$$\dot{R}_{0,\theta_*}(h) := \lim_{c \to 0} \frac{R_0(\theta_* + ch) - R_0(\theta_*)}{c} = \frac{d}{dc} R_0(\theta_* + ch)\bigg|_{c=0}$$

exists and is finite, and furthermore, the functional $\dot{R}_{0,\theta_*} : \mathcal{H}(\theta_*) \to \mathbb{R}$ is linear. The Gâteaux derivative describes the rate at which the risk functional R_0 changes in value when making an infinitesimal shift away from θ_* in the direction h. The key observation we use is that, since θ_0 is an optimizer of R_0, the Gâteaux derivative $\dot{R}_{0,\theta_0}(h)$ of R_0 at θ_0 must be zero in any direction h, that is, $\dot{R}_{0,\theta_0}(h) = 0$ for each $h \in \mathcal{H}(\theta_0)$. Thus, under $H_0 : \theta_0 = \theta_*$, it must also be that $\dot{R}_{0,\theta_*}(h) = 0$ for each $h \in \mathcal{H}(\theta_*)$. Conversely, if H_1 is instead true, there must exist some function $h_* \in \mathcal{H}(\theta_*)$ such that $\dot{R}_{0,\theta_*}(h_*) \neq 0$. To test H_0 against H_1, we assess the existence of such a direction h_*.

Formally, our objective can be restated as determining whether the Gâteaux derivative of R_0 at θ_* in the steepest direction is zero, that is, we note that the null hypothesis H_0 can

be reframed as

$$H_0 : \sup_{h \in \mathcal{H}(\theta_*)} |\dot{R}_{0,\theta_*}(h)| = 0 \ .$$

If the function class Θ — and therefore $\mathcal{H}(\theta_*)$ as well — is rich, it may not be feasible to determine if $\dot{R}_{0,\theta_*}(h) \neq 0$ for any direction $h \in \mathcal{H}(\theta_*)$. We may instead consider a subclass $\mathcal{H} \subseteq \mathcal{H}(\theta_*)$ of directions to investigate, and then assess whether there is empirical evidence to reject the restricted null hypothesis

$$H_{0,r} : \sup_{h \in \mathcal{H}} |\dot{R}_{0,\theta_*}(h)| = 0 \ .$$

Of course, if $H_{0,r}$ is not true, then neither is H_0, but the converse statement does not hold. Thus, a calibrated test of $H_{0,r}$ against its complement will generally constitute a conservative test of H_0 against its complement. We refer to our method as the *restricted score test* because we aim assess $H_{0,r}$ against its complement $H_{1,r}$, that is, to determine if the Gâteaux derivative of greatest magnitude over the restricted space $\mathcal{H}$ is zero or not.

Our approach to assessing $H_{0,r}$ consists of measuring the aggregate 'size' of the collection of Gâteaux derivatives evaluated at θ_* in each direction $h \in \mathcal{H}$. To do so, we may select any norm Ω defined on the vector space $\ell^\infty(\mathcal{H})$ of bounded real-valued functionals on $\mathcal{H}$, and then use $\Omega(\dot{R}_{0,\theta_*})$ as a measure of departure from $H_{0,r}$. Such an approach is valid because if $H_{0,r}$ holds, $\Omega(\dot{R}_{0,\theta_*}) = 0$. We later show that, under appropriate conditions on $\mathcal{H}$, we can construct an estimator $\dot{R}_{n,\theta_*}$ of $\dot{R}_{0,\theta_*}$ such that, as a random element in $\ell^\infty(\mathcal{H})$, the normalized process $\{n^{1/2}[\dot{R}_{n,\theta_*}(h) - \dot{R}_{0,\theta_*}(h)] : h \in \mathcal{H}\}$ converges weakly to a tight mean-zero Gaussian process $\mathbb{G} := \{\mathbb{G}(h) : h \in \mathcal{H}\}$ relative to the supremum norm. Because $\dot{R}_{0,\theta_*}$ is the zero element whenever $\theta_0 = \theta_*$, we may then test $H_{0,r}$ using the test statistic $\Omega(n^{1/2}\dot{R}_{n,\theta_*})$, which has a tractable limiting distribution that can be approximated using resampling techniques, as we demonstrate later.

We propose to construct confidence regions for θ_0 by inverting our restricted score test. Let $\Psi := \{\psi_u : u \in \mathcal{U}\}$ denote a collection of real-valued functionals on Θ indexed by some set $\mathcal{U}$. We define $\mathcal{C}_n := \{\theta \in \Theta :$ we fail to reject $\theta_0 = \theta$ against $\theta_0 \neq \theta$ based on $Z_1, Z_2, \ldots, Z_n\}$

as the set of functions in Θ compatible with the available data and set $\Psi_n(u) := \{\psi_u(\theta) : \theta \in \mathcal{C}_n\}$. Later, we show that a simultaneous confidence region for any smooth functional $\psi_u(\theta_0)$ is given by

$$(\inf \Psi_n(u),\ \sup \Psi_n(u)) \ .$$

In particular, we can set $\mathcal{U}$ to be some subset $\mathcal{O}$ of the domain of θ_0 and take $\psi_u : \theta \mapsto \theta(u)$ to be the evaluation functional at u to obtain a simultaneous confidence set for θ_0 over $\mathcal{O}$.

4.3 Estimation of the risk functional derivative

4.3.1 Uniform asymptotic linearity of the derivative estimator

Having outlined a sketch of the proposed framework for inference, we now scrutinize estimation of the Gâteaux derivative of R_0, which serves as a primary building block of our procedure. To begin, we require that we have at our disposal, for each $h \in \mathcal{H}$, an asymptotically linear estimator $\dot{R}_{n,\theta_*}(h)$ of $\dot{R}_{0,\theta_*}(h)$, in the sense that

$$\dot{R}_{n,\theta_*}(h) - \dot{R}_{0,\theta_*}(h) = \frac{1}{n}\sum_{i=1}^{n} \phi_{P_0,\theta_*}(Z_i; h) + r_{n,\theta_*}(h) \ , \tag{4.4}$$

where $\mathbb{E}_0[\phi_{P_0,\theta_*}(Z; h)] = 0$, $\mathbb{E}_0[\phi_{P_0,\theta_*}(Z; h)^2] < \infty$, and $r_{n,\theta_*}(h) = o_P(n^{-1/2})$. The function $z \mapsto \phi_{P_0,\theta_*}(z; h)$ is referred to as the influence function of $\dot{R}_{n,\theta_*}(h)$. In many settings, such an estimator is readily available. For instance, if the model space $\mathcal{M}$ includes the empirical distribution P_n and the functional $P \mapsto \dot{R}_{P,\theta_*}(h)$ is Hadamard differentiable with respect to the supremum norm (see, e.g., Chapter 20 of van der Vaart, 2000), the plug-in estimator $\dot{R}_{n,\theta_*}(h) := \dot{R}_{P_n,\theta_*}(h)$ will be asymptotically linear with influence function defined pointwise as

$$\phi_{P_0,\theta_*}(z; h) = \frac{d}{d\epsilon}\dot{R}_{P_\epsilon,\theta_*}(h)\bigg|_{\epsilon=0} ,$$

where $P_\epsilon := (1-\epsilon)P_0 + \epsilon\delta_z$ and δ_z is a degenerate distribution on $\{z\}$. Hadamard differentiability typically holds in simple examples, such as when the risk functional has the form (4.1) with loss ℓ_P not depending on P. In other problems, the plug-in estimator $\dot{R}_{P_n,\theta_*}(h)$ may fail

to even be defined — this often occurs when the Gâteaux derivative functional depends on local features of the underlying distribution (e.g., a density or conditional mean function). Provided $P \mapsto \dot{R}_{P,\theta_*}(h)$ is pathwise differentiable relative to the model $\mathcal{M}$ (see, e.g., Bickel et al., 1998), more broadly applicable strategies for estimating $\dot{R}_{0,\theta_*}(h)$ exist. For example, if $\widehat{P}_n \in \mathcal{M}$ is a consistent estimator of P_0, possibly obtained via flexible learning strategies (e.g., machine learning), then the one-step debiased estimator

$$\dot{R}_{n,\theta_*}(h) := \dot{R}_{\widehat{P}_n,\theta_*}(h) + \frac{1}{n}\sum_{i=1}^{n} \phi_{\widehat{P}_n,\theta_*}(Z_i; h)$$

satisfies (4.4) under certain regularity conditions, provided $z \mapsto \phi_{P,\theta_*}(z; h)$ is taken to be any gradient of the pathwise derivative of $P \mapsto \dot{R}_{P,\theta_*}(h)$ (Pfanzagl, 1982). Alternative constructions, such as targeted minimum loss-based estimation (see, e.g., van der Laan and Rose, 2011), also exist.

By the central limit theorem, the asymptotic representation (4.4) suffices to establish that, for any finite subset $\mathcal{H}_0 \subset \mathcal{H}$, $\{n^{1/2}[\dot{R}_{n,\theta_*}(h) - \dot{R}_{0,\theta_*}(h)] : h \in \mathcal{H}_0\}$ converges in distribution to a mean-zero Gaussian random vector. This does not readily extend to an infinite set $\mathcal{H}_0$ — or indeed, $\mathcal{H}$ itself — without imposing stronger requirements on $\dot{R}_{n,\theta_*}$, as is needed in our proposal. The following lemma provides additional conditions on $\dot{R}_{n,\theta_*}$ under which joint asymptotic normality of $n^{1/2}[\dot{R}_{n,\theta_*}(h) - \dot{R}_{0,\theta_*}(h)]$ holds over an infinite set $\mathcal{H}_0$.

Lemma 4.1. *If (i)* $\{z \mapsto \phi_{0,\theta_*}(z; h) : h \in \mathcal{H}\}$ *is a* P_0*-Donsker class, and (ii)* $\sup_{h\in\mathcal{H}} |r_n(h)| = o_P(n^{-1/2})$*, then, as an element of* $\ell^\infty(\mathcal{H})$*,* $\{n^{1/2}[\dot{R}_{n,\theta_*}(h) - \dot{R}_{0,\theta_*}(h)] : h \in \mathcal{H}\}$ *converges weakly to a tight mean-zero Gaussian process* $\mathbb{G}$ *with covariance function* $\Sigma : (h_1, h_2) \mapsto \mathbb{E}_0\left[\phi_{P_0,\theta_*}(Z; h_1)\phi_{P_0,\theta_*}(Z; h_2)\right]$ *relative to the supremum norm.*

Both conditions constrain the complexity of $\mathcal{H}$. Condition (i) often holds, for example, if $\mathcal{H}$ is itself a P_0-Donsker class, as implied by Theorem 2.10.6 of van der Vaart and Wellner (1996). Condition (ii) requires that the asymptotic linearity of $\dot{R}_{n,\theta_*}(h)$ hold uniformly for $h \in \mathcal{H}$. It is trivially satisfied irrespective of $\mathcal{H}$ if, for example, $P_n \in \mathcal{M}$, $P \mapsto \dot{R}_{P,\theta_*}(h)$ is a linear functional, and the plug-in estimator $\dot{R}_{P_n,\theta_*}(h)$ is used.

4.3.2 Working examples

Example 1: nonparametric mean regression

We first consider the setting of nonparametric mean regression, in which θ_0 is the conditional mean function $x \mapsto \mathbb{E}_0(Y \mid X = x)$, which is also expressed as a population risk minimizer in (4.2). For a fixed direction h, the Gâteaux derivative of $\theta \mapsto R_P(\theta)$ at $\theta = \theta_*$ takes the form $\dot{R}_{P,\theta_*}(h) = \mathbb{E}_P\{[Y - \theta_*(X)]h(X)\}$. In particular, this suggests that, in this problem, the score test can be interpreted as examining the orthogonality of the residual calculated under the null hypothesis $H_0 : \theta_* = \theta_0$ to all functions $h \in \mathcal{H}$.

The fact that $P \mapsto \dot{R}_{P,\theta_*}(h)$ is a linear functional defined at the empirical distribution P_n suggests the use of the plug-in estimator

$$\dot{R}_{n,\theta_*}(h) := \frac{1}{n}\sum_{i=1}^{n}[Y_i - \theta_*(X_i)]\,h(X_i)$$

of $\dot{R}_{0,\theta_*}(h)$. This plug-in estimator is in fact unbiased and asymptotically linear with influence function

$$z = (x, y) \mapsto \phi_{P_0,\theta_*}(z; h) := [y - \theta_*(x)]h(x) - \dot{R}_{0,\theta_*}(h)\ .$$

Since this influence function is the efficient influence function of $P \mapsto \dot{R}_{P,\theta_*}(h)$ relative to a nonparametric model, $\dot{R}_{n,\theta_*}(h)$ is also nonparametric efficient. The remainder term $r_n(h)$ in (4.4) is exactly zero, so the conditions of Lemma 4.1 are often satisfied as long as $\mathcal{H}$ satisfies a Donsker condition. Uniform convergence of $\dot{R}_n(\theta_*; h)$ is thus achieved under relatively weak conditions.

Example 2: partially additive mean regression

We now consider the setting of a partially additive mean model. This example is more involved than the previous example because the risk functional $P \mapsto R_P$ is nonlinear and depends on P via an unknown function-valued nuisance parameter f_P. It is possible to verify that this nuisance parameter can be expressed as $f_P : w \mapsto \mathbb{E}_P[Y - \theta_P(X) \mid W = w]$. Coupled

with (4.3), this fact implies that θ_P minimizes the population P-risk functional $R_{P,\theta_*}(h) := \mathbb{E}_P\left[Y - \mu_{Y,P}(W) - \{\theta_*(X) - \mathbb{E}_P\left[\theta_*(X) \,|\, W\right]\}\right]^2$, where we define $\mu_{Y,P} : w \mapsto \mathbb{E}_P\left(Y \,|\, W = w\right)$. The Gâteaux derivative of this risk functional takes the form

$$\dot{R}_{P,\theta_*}(h) = \mathbb{E}_P\left\{[Y - \mu_{Y,P}(W) - \theta_*(X) + \mu_{\theta_*,P}(W)][h(X) - \mu_{h,P}(W)]\right\}, \tag{4.5}$$

where we define $\mu_{g,P} : w \mapsto \mathbb{E}_P\left[g(X) \,|\, W = w\right]$ for each function $g : \mathbb{R} \to \mathbb{R}$ for which this moment exists.

Obtaining an estimator of $\dot{R}_{0,\theta_*}(h)$ is more challenging than in the previous example, as we need to estimate the nuisance parameters μ_{Y,P_0}, μ_{h,P_0} and μ_{θ_*,P_0}. Suppose that we have constructed consistent estimators μ_{n,Y,P_0}, μ_{n,h,P_0} and μ_{n,θ_*,P_0} of these nuisance functions using a nonparametric estimation procedure, such as artificial neural networks (Barron, 1989), the highly adaptive lasso (Benkeser and van der Laan, 2016), or the Super Learner (van der Laan et al., 2007). In Result C.1 in Appendix C.1, we show that the resulting plug-in estimator,

$$\dot{R}_{n,\theta_*}(h) := \frac{1}{n}\sum_{i=1}^{n}[Y_i - \mu_{n,Y,P_0}(W_i) + \mu_{n,\theta_*,P_0}(W_i) - \theta_*(X_i)][h(X_i) - \mu_{n,h,P_0}(W_i)], \tag{4.6}$$

is asymptotically linear with influence function

$$z = (w, x, y) \mapsto \phi_{P_0,\theta_*}(z; h) := \{y - \mu_{Y,P_0}(w) + \mu_{\theta_*,P_0}(w) - \theta_*(x)\}\{h(x) - \mu_{h,P_0}(w)\} - \dot{R}_{0,\theta_*}(h) \tag{4.7}$$

under rate and complexity conditions on nuisance estimators μ_{n,Y,P_0}, μ_{n,h,P_0} and μ_{n,θ_*,P_0}. Again, since this influence function is the efficient influence function $P \mapsto \dot{R}_{P,\theta_*}(h)$ relative to a nonparametric model, $\dot{R}_{n,\theta_*}(h)$ is also nonparametric efficient.

While the asymptotic linearity of $\dot{R}_{n,\theta_*}(h)$ at a fixed h can be readily established, uniform asymptotic linearity over $\mathcal{H}$ is more difficult to achieve and requires stronger conditions, stated explicitly in Appendix C.1. In particular, uniform control of the remainder from the linear representation of $\dot{R}_{n,\theta_*}(h)$ (condition ii of Lemma 1) requires consistent estimation of the nuisance function μ_{h,P_0} uniformly over $\mathcal{H}$. For large $\mathcal{H}$, this may be a difficult feat. Additionally, computational difficulties may arise since computing an estimate of μ_{h,P_0} separately

for each $h \in \mathcal{H}$ can be unfeasible unless a convenient parametrization of $\mathcal{H}$ is available. We provide such a construction in Section 4.5.

4.4 Properties of the restricted score test

4.4.1 Limiting distribution of the test statistic

We now describe the construction of our restricted score test based on an estimator $\dot{R}_{n,\theta_*}$ satisfying the conditions outlined in Section 4.3 and establish its large-sample properties.

Whenever $\theta_0 = \theta_*$, the test statistic $T_n := \Omega(n^{1/2}\dot{R}_{n,\theta_*})$ converges in distribution to $\Omega(\mathbb{G})$ for any norm Ω on $\ell^\infty(\mathcal{H})$. We wish to compute the distribution function of $\Omega(\mathbb{G})$ in order to obtain an approximate p-value $\rho(t)$ based on an arbitrary realization t of T_n. However, since the limiting distribution of T_n is generally not available in closed form, we will use resampling techniques.

We propose a multiplier bootstrap procedure that leverages the asymptotic linearity of $\dot{R}_{n,\theta_*}$ to approximate the p-value $\rho(t)$. For each $m = 1, 2, \ldots, M$, let $\xi_{m,1}, \xi_{m,2}, \ldots, \xi_{m,n}$ be a random sample of independent and identically distributed random variables (also independent of $Z_1, Z_2, \ldots, Z_n$) with mean zero, unit variance and finite moment of order $2 + \omega$ for some $\omega > 0$. For instance, these could be taken as a random sample of Rademacher or standard normal random variables. Defining the bootstrapped mapping $\dot{R}_{m,n,\theta_*} : h \mapsto \frac{1}{n}\sum_{i=1}^n \phi_{n,\theta_*}(Z_i, h)\xi_{m,i}$, where ϕ_{n,θ_*} a consistent estimator of ϕ_{P_0,θ_*}, we construct the bootstrapped test statistic

$$T_{m,n} := \Omega\left(n^{1/2}\dot{R}_{m,n,\theta_*}\right) . \tag{4.8}$$

Under suitable regularity conditions, this statistic converges weakly to $\Omega(\mathbb{G})$, and so,

$$\rho_{M,n}(t) := \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}\left(T_{m,n} > t\right)$$

serves as an approximation to $\rho(t)$ for n and M large. This result is stated formally below, in Theorem 4.1. Below, we suppose that the influence function ϕ_{P_0,θ_*} depends on P_0 only

through some nuisance parameter $f_0 \in \mathcal{F}$, where $\mathcal{F}$ is a vector space endowed with some norm $\|\cdot\|_{\mathcal{F}}$. With some abuse of notation, for any candidate nuisance $f \in \mathcal{F}$, we denote by ϕ_{f,θ_*} the influence function corresponding to nuisance value f. We note then, in particular, that $\phi_{f_0,\theta_*} = \phi_{P_0,\theta_*}$. We consider this representation of ϕ_{P_0,θ_*} show that implementing our procedure may not require estimating the entire distribution P_0 but only some summary of P_0 (e.g., a mean value or conditional mean function under P_0). There are no restrictions on this summary, as we may also take f_0 to be the density function of P_0, if appropriate. We suppose that we have access to an estimator f_n of f_0 based on $Z_1, Z_2, \ldots, Z_n$.

Theorem 4.1. *Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent and identically distributed random variables with mean zero, variance one and finite raw moment of order $2+\omega$ for some $\omega > 0$, and also independent of $Z_1, Z_2, \ldots, Z_n$. Suppose that, for some $\delta > 0$, the class $\Phi_\delta := \{z \mapsto \phi_{f,\theta_*}(z;h) - \phi_{f_0,\theta_*}(z;h) : h \in \mathcal{H}, f \in \mathcal{F}, \|f - f_0\|_{\mathcal{F}} < \delta\}$ is P_0-Donsker and has a finite envelope function. Then, provided that $\|f_n - f_0\|_{\mathcal{F}} = o_P(1)$ and that*

$$\sup_{h \in \mathcal{H}} \int [\phi_{f,\theta_*}(z;h) - \phi_{f_0,\theta_*}(z;h)]^2 dP_0(z) \longrightarrow 0$$

as $\|f - f_0\|_{\mathcal{F}} \to 0$, $\left\{n^{-1/2} \sum_{i=1}^n \xi_i \phi_{f_n,\theta_*}(Z_i;h) : h \in \mathcal{H}\right\}$ converges weakly to $\mathbb{G}$ relative to the supremum norm as an element of $\ell^\infty(\mathcal{H})$ conditional upon the sample paths $Z_1, Z_2, \ldots, Z_n$, in outer probability.

We recall that the Gaussian process $\mathbb{G}$ was explicitly defined in Lemma 4.1. In view of this result, the resampling-based strategy described above can be used to approximate the limiting distribution of the test statistic under $H_0 : \theta_0 = \theta_*$. However, given that the limiting distributions of $\Omega(n^{1/2}\dot{R}_{n,\theta_*})$ and $\Omega(n^{1/2}\dot{R}_{n,\theta_0})$ coincide under the null hypothesis, we could instead obtain and use an approximation to the latter. This strategy has the key advantage that the same bootstrap samples can be used to test $H_0 : \theta_0 = \theta_*$ for *any* value of θ_*. This can lead to large gains in computational efficiency when constructing confidence bands, as we discuss in Section 4.4.4. Using the tools used to prove Theorem 4.1, it is possible to derive a slight modification of the result in which θ_* is replaced by a consistent estimator θ_n.

4.4.2 Selection of the class of directions and norm

While our proposed test of $H_{0,r}$ — and thus of H_0 — achieves nominal type I error control with any norm Ω and any sufficiently small class $\mathcal{H}$, statistical power is influenced by these selections. The optimal norm and class of directions as well as the sensitivity of power to their selection depend on properties of the underlying data-generating mechanism P_0.

We first discuss selection of $\mathcal{H}$. To do so, we examine the local asymptotic power of the restricted score test when $\mathcal{H}$ is a singleton set containing a fixed direction h. Suppose that $Z_1, Z_2, \ldots, Z_n$ are generated from a distribution $P_{0,n}$ such that the Gâteaux derivative under $P_{0,n}$ is $n^{-1/2}t_h$ for $t_h \in \mathbb{R}$. Suppose also that $\dot{R}_{n,\theta_*}(h)$ is locally regular in the sense that

$$\dot{R}_{n,\theta_*}(h) - n^{-1/2}t_h = \frac{1}{n}\sum_{i=1}^{n} \phi_{0,\theta_0}(Z_i; h) + r_n(h) \ ,$$

where the influence function ϕ_{0,θ_0} does not depend on n, and $n^{1/2}r_n(h)$ converges to zero in probability under sampling from $P_{0,n}$. It can be shown through an application of the Neyman-Pearson lemma that the most powerful test based on $\dot{R}_{n,theta_*}(h)$ of the null hypothesis that $\dot{R}_{0,\theta_*}(h) = 0$ rejects the null when the statistic $T_n^* := \dot{R}^2_{n,\theta_*}(h)/\mathbb{E}_0\left[\phi^2_{0,\theta_0}(Z; h)\right]$ is larger than the $(1-\alpha)$-quantile of the χ^2_1 distribution, and that T_n^* approximately follows a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $t_h^2/\mathbb{E}_0\left[\phi^2_{0,\theta_0}(Z; h)\right]$ for large n. Hence, the local asymptotic power is determined by the ratio of the Gâteaux derivative to the asymptotic variance of $n^{1/2}[\dot{R}_{n,\theta_0}(h) - \dot{R}_{0,\theta_0}(h)]$,

$$\frac{\dot{R}^2_{0,\theta_*}(h)}{\mathbb{E}_0\left[\phi^2_{0,\theta_0}(Z; h)\right]} \ . \tag{4.9}$$

Thus, if we perform a test by taking $\mathcal{H}$ to be a set containing only a single fixed direction, an optimal direction would be any maximizer h_0 of (4.9) over $h \in \mathcal{H}(\theta_*)$. It is therefore reasonable to seek to select $\mathcal{H}$ as a small set of functions that contains a good approximation of h_0. Because h_0 is unknown and will typically depend on P_0, compelling approach consists of characterizing h_0 analytically and then taking $\mathcal{H}$ to be a class of smooth functions that contains an estimate of h_0. We provide explicit details regarding the construction of $\mathcal{H}$ in Section 4.5.

We now provide natural examples of the norm Ω. Let $V : \mathcal{H} \to [0, \infty)$ be a non-negative weight functional. We first consider the weighted supremum norm

$$a \mapsto \Omega_\infty(a) := \sup_{h \in \mathcal{H}} V(h)|a(h)| \ ,$$

leading to the test statistic $\Omega_\infty(n^{1/2}\dot{R}_{n,\theta_*}) = n^{1/2} \sup_{h \in \mathcal{H}} V(h)|\dot{R}_{n,\theta_*}(h)|$. With the choice $V \equiv 1$, this simply evaluates $n^{1/2}\dot{R}_{n,\theta_*}$ at the direction of greatest estimated change. Under the alternative hypothesis, the largest estimated Gâteaux derivative value does not necessarily provide the greatest evidence in favor of the alternative because of the variability of the derivative estimator. This motivates us to instead weight the Gâteaux derivative by the reciprocal of the standard deviation implied by the influence function of the derivative estimator, thus setting $V = V_0$ with $V_0(h) := \{\mathbb{E}_0 [\phi_{P_0,\theta_0}(Z; h)^2]\}^{-1/2}$. We note that, when Θ is finite-dimensional and $\mathcal{H} = \mathcal{H}(\theta_*)$, the statistic resulting from use of the variance-weighted supremum norm is precisely equivalent to the original score statistic proposed in Rao (1948). The variance-weighted supremum norm can thus be viewed as a natural generalization of the standard parametric score test to the infinite-dimensional setting. We note that our suggested choice of weight V_0 depends on P_0, and so, in practice, we must use an estimator V_n of V_0. It can be seen through an application of Slutsky's theorem that, as long as V_n is uniformly consistent, in the sense that $\sup_{h \in \mathcal{H}} |V_n(h) - V_0(h)| = o_P(1)$, our theoretical results remain valid.

As an alternative norm, we also consider a weighted L_2 norm over $\mathcal{H}$. Let Q be a measure on the Borel σ-algebra generated by $\mathcal{H}$. We define the weighted L_2 norm as

$$a \mapsto \Omega_2(a) := \left\{ \int_{\mathcal{H}} [V(h)a(h)]^2 dQ(h) \right\}^{1/2}$$

and consider the test statistic $\Omega_2(n^{1/2}\dot{R}_{n,\theta_*})$. This statistic involves weighted averaging of the Gâteaux derivative corresponding to each direction $h \in \mathcal{H}$. Similarly as with the supremum norm, we may wish to set $V = V_0$ in order to place more weight on directions for which we can estimate the Gâteaux derivative with greater precision.

To understand the influence of the choice of Ω on power, we draw intuition from literature on simultaneous testing in the high-dimensional setting (e.g., Cai et al., 2014). In settings

where the signal is dense, in the sense that the Gâteaux derivative is small but nonzero in many directions in $\mathcal{H}$, good performance is expected from the L_2 norm but not from the supremum norm. Conversely, when the signal is sparse, in the sense that the Gâteaux derivative is large in relatively few directions and zero elsewhere, the supremum norm is expected to yield better performance than the L_2 norm.

4.4.3 Extension to data-dependent classes of directions

So far, we have considered $\mathcal{H}$ to be a fixed class. In practice, it can be difficult to select $\mathcal{H}$ a priori, and we may want to instead select the class of directions in a data-driven manner. To be applicable in such cases, our theoretical results must allow the fixed class $\mathcal{H}$ to be replaced by a stochastic (data-dependent) sequence of classes $\mathcal{H}_n = \mathcal{H}_n(Z_1, Z_2, \ldots, Z_n)$. The following theorem indicates that if $\mathcal{H}_n$ converges to a fixed class $\mathcal{H}$ in an appropriate sense, then for the two choices of norm we have considered, $\Omega(n^{1/2}\dot{R}_{n,\theta_*})$ also converges weakly to $\Omega(\mathbb{G})$, where $\mathbb{G}$ is the same Gaussian process defined in Lemma 4.1. Below, to simplify the notation, we fix θ_* and denote by ϕ_h the function $z \mapsto \phi_{P_0,\theta_*}(z;h)$.

Theorem 4.2. *Suppose that there exists a function class $\bar{\mathcal{H}}$ such that $\{\phi_h : h \in \bar{\mathcal{H}}\}$ is a P_0-Donsker class with a finite and square-integrable envelope, and that $\mathcal{H}_n \cup \mathcal{H} \subseteq \bar{\mathcal{H}}$ with P_0-probability one. Suppose also that $H_0 : \theta_* = \theta_0$ holds.*

(a) If $\{\phi_h : h \in \mathcal{H}_n\}$ converges to $\{\phi_h : h \in \mathcal{H}\}$ in the Hausdorff sense, that is,

$$\max\left\{\sup_{h_1\in\mathcal{H}}\inf_{h_2\in\mathcal{H}_n}\int[\phi_{h_1}(z)-\phi_{h_2}(z)]^2dP_0(z), \sup_{h_2\in\mathcal{H}_n}\inf_{h_1\in\mathcal{H}}\int[\phi_{h_1}(z)-\phi_{h_2}(z)]^2dP_0(z)\right\} = o_P(1)\ , \tag{4.10}$$

then $\sup_{h\in\mathcal{H}_n} n^{1/2}\int\phi_h(z)d(P_n-P_0)(z)$ converges in distribution to $\sup_{h\in\mathcal{H}}\mathbb{G}(\phi_h)$.

(b) Let $\mathcal{B}(\bar{\mathcal{H}})$ denote the Borel σ-algebra, and let $\bar{Q}$ be a measure on $\mathcal{B}(\bar{\mathcal{H}})$. If

$$\bar{Q}\left(\{\mathcal{H}\cup\mathcal{H}_n\}\setminus\{\mathcal{H}\cap\mathcal{H}_n\}\right) = o_P(1)\ , \tag{4.11}$$

then $\int_{\mathcal{H}_n} n\{\int\phi_h(z)d(P_n-P_0)(z)\}^2d\bar{Q}(h)$ converges in distribution to $\int_{\mathcal{H}}\int\{\mathbb{G}(\phi_h)\}^2d\bar{Q}(h)$.

Theorem 4.2 can be applied to conclude that $\Omega\left(n^{1/2}\dot{R}_{n,\theta_*}\right)$ converges weakly to $\Omega(\mathbb{G})$ if $\dot{R}_{n,\theta_*}(h)$ is uniformly asymptotically linear for $h \in \bar{\mathcal{H}}$, with $\bar{\mathcal{H}}$ defined in the theorem statement. For convergence of the supremum norm, Theorem 4.2 requires that for any direction $h_1 \in \mathcal{H}$, there exists a direction $h_2 \in \mathcal{H}_n$ such that the expected squared difference between the influence functions for the Gâteaux derivative estimator corresponding to directions h_1 and h_2 converges in probability to zero, and similarly for any direction $h_2 \in \mathcal{H}_n$. For convergence of the L_2 norm, we require that the measure of the difference between the union and intersection of $\mathcal{H}$ and $\mathcal{H}_n$ converges in probability to zero. While we have not shown that the multiplier bootstrap scheme described in Section 4.4.1 also provides a valid approximation of the sampling distribution of $\Omega(n^{1/2}\dot{R}_{n,\theta_*})$ when the class of directions is data-dependent, we expect that the multiplier bootstrap remains valid under suitable regularity conditions.

4.4.4 Construction of confidence regions

By taking advantage of the relationship between hypothesis tests and confidence regions, we can invert the proposed score test to obtain simultaneous confidence sets for summaries of θ_0. Let $\Psi := \{\psi_u : u \in \mathcal{U}\}$ denote a collection of real-valued functionals defined on Θ indexed by some set $\mathcal{U}$. We wish to construct simultaneous confidence intervals for elements of $\Psi(\theta_0) := \{\psi_u(\theta_0) : u \in \mathcal{U}\}$.

As before, we take $\mathcal{C}_n := \{\theta \in \Theta : \text{we fail to reject } \theta_0 = \theta \text{ against } \theta_0 \neq \theta \text{ based on } Z_1, Z_2, \ldots, Z_n\}$ to denote the set of null parameter values that the restricted score test fails to reject. If the test achieves the nominal type I error rate α, θ_0 belongs to $\mathcal{C}_n$ with probability tending to $1 - \alpha$ in the sense that

$$P_0\left(\theta_0 \in \mathcal{C}_n\right) = P_0\left(\text{we fail to reject } \theta_0 = \theta \text{ against } \theta_0 \neq \theta \text{ based on } Z_1, Z_2, \ldots, Z_n\right) \longrightarrow 1 - \alpha$$

as sample size n tends to infinity whenever $H_0 : \theta_0 = \theta$ is true. Thus, $\mathcal{C}_n$ is a $100(1-\alpha)\%$ confidence region for θ_0. The random region $\mathcal{C}_n$ can be interpreted as the collection of parameter values θ that are consistent with the observed data. To obtain a confidence region for $\Psi(\theta_0)$, for each $u \in \mathcal{U}$, we find the largest and smallest value of $\psi_u(\theta)$ that can be obtained

for $\theta \in \mathcal{C}_n$; setting $\Psi_{n,u} := \{\psi_u(\theta) : \theta \in \mathcal{C}_n\}$, we construct the set

$$\mathcal{C}_n(u) := (\inf \Psi_{n,u},\ \sup \Psi_{n,u}) \ .$$

To see that $\{\mathcal{C}_n(u) : u \in \mathcal{U}\}$ is in fact a simultaneous confidence region, we note that if $\psi_{u'}(\theta_0) \notin \mathcal{C}_n(u')$ for some $u' \in \mathcal{U}$, then it is necessarily the case that $\theta_0 \notin \mathcal{C}_n$. Thus, we have that

$$\liminf_{n\to\infty} \mathbb{P}_0 \left\{\psi_u(\theta_0) \in \mathcal{C}_n(u) \text{ for all } u \in \mathcal{U}\right\} \ \geq \ \liminf_{n\to\infty} \mathbb{P}_0 \left(\theta_0 \in \mathcal{C}_n\right) \ = \ 1 - \alpha \ .$$

In some instances, if Θ is unrestricted, an interval $\mathcal{C}_n(u)$ can be infinitely wide. This can occur when it is possible to construct several non-smooth functions θ such that $\dot{R}_{n,\theta}(h) = 0$ for all h. For instance, in Example 1, this can be achieved by fixing $\theta(X_i) = Y_i$ for $i = 1, 2, \ldots, n$, and letting θ take *any* value elsewhere. Thus, one can select $\theta \in \mathcal{C}_n$ such that the evaluation $\theta(x_0)$ of θ at a point x_0 where no data are observed is arbitrarily large or small. This difficulty is avoided if Θ is a class of smooth functions, although determining such a class *a priori* may be difficult in practice. If we could obtain a consistent estimator of a measure of the smoothness of θ_0, we could construct a confidence band for a class containing functions no smoother than the smoothness level prescribed by our estimate. However, it can be challenging to consistently estimate the smoothness of an unknown function. In our implementation, we use a simple plug-in estimator of the smoothness and leave the development of a more rigorous approach as future work.

Our proposal is particularly useful for obtaining simultaneous confidence intervals for the evaluation functional $\theta \mapsto \theta(x)$ for arbitrary x, as we are able to immediately obtain a $100(1-\alpha)\%$ confidence band for θ_0. When only a finite collection of functionals is of interest, however, our proposed intervals may be conservative. If the functionals of interest are pathwise differentiable, we recommend instead estimating each quantity using an efficient estimator (as described in, e.g., Pfanzagl, 1982 and van der Laan and Rose, 2011) to obtain asymptotically calibrated intervals.

4.5 Implementation and practical considerations

We now describe our implementation of the restricted score test and discuss its use in the partially additive mean regression example.

4.5.1 Construction of $\mathcal{H}$

For a positive semidefinite kernel function K, let $\mathcal{S}_K$ denote its unique reproducing kernel Hilbert space (RKHS), endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{S}_K}$. We construct $\mathcal{H}$ to be a subspace of $\mathcal{S}_K$.

We consider the eigen-decomposition of K given by

$$(z_1, z_2) \mapsto K(z_1, z_2) = \sum_{j=1}^{\infty} \kappa_j \eta_j(z_1) \eta_j(z_2)$$

with eigenfunctions $\{\eta_1, \eta_2, \ldots\}$ orthogonal with respect to $\langle \cdot, \cdot \rangle_{\mathcal{S}_K}$ and eigenvalues $0 \leq \kappa_1 < \kappa_2 < \ldots$. Each function $g \in \mathcal{S}_K$ can be expressed as a linear combination $z \mapsto \sum_{j=1}^{\infty} a_j \eta_j(z)$ of eigenfunctions, where $a_1, a_2, \ldots$ are real-valued coefficients. The smoothness of g can be measured by the RKHS norm as

$$J(g) := \langle g, g \rangle_{\mathcal{S}_K} = \sum_{j=1}^{\infty} \frac{a_j^2}{\kappa_j}$$

with higher values of $J(g)$ corresponding to lesser smoothness. The test statistics we consider are based on the ratio of the Gâteaux derivative estimates to their standard errors and do not depend on the scale of the direction function h. Rather, the performance of our test is determined by the *shape* of the direction function. We therefore require a scale-free measure of smoothness, for which we use a scaled version of the RKHS norm $J(h)V_0^2(h)$, where we recall that $V_0(h) := \{\mathbb{E}_0 [\phi_{P_0,\theta_0}(Z; h)^2]\}^{-1/2}$ is the reciprocal of the asymptotic standard deviation of $n^{1/2}[\dot{R}_{n,\theta_0}(h) - \dot{R}_{0,\theta}(h)]$. While other scale-free measurements of smoothness could alternatively be used, we will see that this particular choice leads to computational benefits. We consider as $\mathcal{H}$ a subset of functions in $\mathcal{S}_K$ with bounded smoothness, namely

$$\mathcal{H}_\gamma := \left\{ h = \textstyle\sum_{j=1}^{\infty} a_j \eta_j : a_1, a_2, \ldots \in \mathbb{R}, J(h)V_0^2(h) \leq \gamma \right\}$$

for some tuning value $\gamma > 0$. For computational ease, we truncate the eigenbasis at some large level d.

In some instances, the kernel for an RKHS has eigenfunctions that are known and available in closed form (Wahba, 1990). For instance, consider the second-order Sobolev space on $[0, 1]$, which can be defined as an RKHS endowed with the inner product $(h_1, h_2) \mapsto \langle h_1, h_2 \rangle_{\mathcal{S}_K} = \int_0^1 \ddot{h}_1(z)\ddot{h}_2(z)dz$, where $\ddot{h}$ denotes the second derivative of any given function h. The eigenfunctions and eigenvalues for the kernel are

$$\eta_{2j-1} : z \mapsto \sqrt{2}\cos(2\pi j z), \quad \eta_{2j} : z \mapsto \sqrt{2}\sin(2\pi j z), \quad \kappa_{2j-1} = \kappa_{2j} = (2\pi j)^{-4},$$

for $j = 1, 2, \ldots$. When the eigenfunctions are not available in closed form, we can instead use a numerical approximation. For instance, letting $\mathbf{K}$ be an $n \times n$ matrix with $\mathbf{K}_{ij} := K(Z_i, Z_j)$, we could take $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_d \in \mathbb{R}^n$ as the leading eigenvectors of $\mathbf{K}$.

For the partially additive mean model, estimation of the Gâteaux derivative requires estimation of the conditional mean μ_{h,P_0} of $h(X)$ given W under P_0. We proceed by first obtaining an estimate μ_{n,η_j,P_0} of the eigenfunction regression function $w \mapsto \mu_{\eta_j,P_0}(w) := \mathbb{E}_0[\eta_j(X) \,|\, W = w]$ and then setting $\mu_{n,\eta,P_0} := \sum_{j=1}^d a_j \mu_{n,\eta_j,P_0}$. The estimate μ_{n,η_j,P_0} may be unreliable for a non-smooth eigenfunction η_j and hence lead to a poor estimate of μ_{h,P_0}. However, by requiring that h is smooth, and thus forcing a_j to be small for non-smooth eigenfunctions, estimates of the non-smooth eigenfunctions should make a relatively small contribution to μ_{n,h,P_0}.

We conclude by discussing selection of the tuning parameter γ. We choose γ so that $\mathcal{H}_\gamma$ contains an approximation of a maximizer h_0 of (4.9). As h_0 may depend on P_0, we instead obtain an estimate h_n of h_0 and take $\gamma = \gamma_n := J(h_n)\, V_n^2(h_n)$, where $V_n(h)$ is an estimate of $V_0(h)$. It can be shown, by an application of the Cauchy-Schwarz inequality, that for the partially additive mean model, if the residual $Y - f_0(W) - \theta_0(X)$ depends neither on X nor W, any maximizer of (4.9) is proportional to $\theta_0 - \theta_*$. We then set $h_0 = \theta_0 - \theta_*$. We use a simple penalization approach to estimate θ_0 and consider candidate estimates of the form $\theta = \sum_{j=1}^d a_j \eta_j$. Let R_n be an estimate of the population risk function R_0. We estimate the

risk for the mean regression function in the partially additive model as

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \mu_{n,Y,P_0}(W_i) - \sum_{j=1}^{d} a_j\left\{\eta_j(X_i) - \mu_{n,\eta_j,P_0}(W_i)\right\}\right]^2.$$

We then estimate θ_0 as $\theta_n := \sum_{j=1}^{d}\hat{a}_j\eta_j$, where

$$(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_d) := \underset{(a_1,a_2,\ldots,a_d)\in\mathbb{R}}{\operatorname{argmin}} R_n\left(\textstyle\sum_{j=1}^{d} a_j\eta_j\right) + \lambda\displaystyle\sum_{j=1}^{d}\frac{a_j^2}{\kappa_j}. \tag{4.12}$$

The penalty term in (4.12) modulates the smoothness the estimator θ_n, where smoothness is measured by the RKHS norm, and a larger value of the tuning parameter $\lambda > 0$ results in a smoother estimate. We select λ using cross-validation, where the performance on the validation set is measured using the empirical risk, and the nuisance parameters upon which the risk depends are still estimated using the full data set. We finally take $h_n := \theta_n - \theta_*$.

4.5.2 Calculation of the test statistic

We now discuss strategies for computing the test statistic. As calculation of the test statistic depends on the form of $\dot{R}_{n,\theta_*}(h)$, we focus specifically on the partially additive mean model.

The estimate $\dot{R}_{n,\theta_*}(h)$ of the Gâteaux derivative (4.5) can be expressed as

$$\dot{R}_{n,\theta_*}(h) = n^{-1}\mathbf{S}(\theta_*)^\top\Gamma\mathbf{a}$$

for each $h \in \mathcal{H}_\gamma$, where $\mathbf{S}(\theta_*)$ is an n-dimensional vector with i^{th} component $\mathbf{S}(\theta_*)_i := Y_i - \mu_{n,Y,P_0}(W_i) - [\theta_*(X_i) - \mu_{n,\theta_*,P_0}(X_i)]$, Γ is an $n \times d$ matrix with $(i,j)^{th}$ entry $\Gamma_{ij} := \eta_j(X_i) - \mu_{n,\eta_j,P_0}(W_i)$, and $\mathbf{a}$ is a d-dimensional vector of coefficients for the eigenbasis. We estimate the variance of the efficient influence function $\phi_{P_0,\theta_0}(Z;h)$ as

$$V_n^{-2}(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{[Y_i - \mu_{n,Y,P_0}(W_i) + \mu_{n,\theta_n,P_0}(W_i) - \theta_n(X_i)][h(X_i) - \mu_{n,h,P_0}(W_i)]\right\}^2.$$

For $h \in \mathcal{H}_\gamma$, we can rewrite the estimate as $V_n^{-2}(h) = \mathbf{a}^\top\mathbf{V}\mathbf{a}$, where $\mathbf{V} = n^{-1}\Gamma^\top[\operatorname{diag}(\mathbf{S}(\theta_n))]^2\Gamma$

.

The inverse variance-weighted supremum norm test statistic $\omega_{\infty,n} := \Omega_\infty(n^{1/2}\dot{R}_{n,\theta_*})$ then takes the form

$$\omega_{\infty,n}^2 = \sup_{\mathbf{a}} \left\{ \frac{n^{-1}[\mathbf{S}(\theta_*)^\top \Gamma \mathbf{a}]^2}{\mathbf{a}^\top \mathbf{V} \mathbf{a}} : \frac{\mathbf{a}^\top \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a}}{\mathbf{a}^\top \mathbf{V} \mathbf{a}} \leq \gamma \right\}, \tag{4.13}$$

where $\boldsymbol{\kappa} := (\kappa_1, \kappa_2, \ldots, \kappa_d)$. A maximizer of the optimization problem in (4.13) can be obtained by solving

$$\sup_{\mathbf{a}} \left\{ n^{-1/2}\mathbf{S}(\theta_*)^\top \Gamma \mathbf{a} : \mathbf{a}^\top \mathbf{V} \mathbf{a} = 1, \mathbf{a}^\top \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a} \leq \gamma \right\}. \tag{4.14}$$

The Karush-Kuhn-Tucker (KKT) conditions for the constrained optimzation problem (4.14) imply that the maximizer of (4.14) is the solution to

$$\operatorname*{argmax}_{\mathbf{a}} \left\{ n^{-1/2}\mathbf{S}(\theta_*)^\top \Gamma \mathbf{a} - \frac{\lambda_2}{2} \left(\mathbf{a}^\top \mathbf{V} \mathbf{a} + \lambda_1 \, \mathbf{a}^\top \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a} \right) \right\}, \tag{4.15}$$

where $\lambda_1, \lambda_2 > 0$ are chosen so that the constraints in (4.14) are satisfied, and the inequalities are tight. With some algebra, we find that the solution $\tilde{\mathbf{a}}_{\lambda_1,\lambda_2}$ to (4.15) is available in closed form as

$$\tilde{\mathbf{a}}_{\lambda_1,\lambda_2} = n^{-1/2}\lambda_2^{-1} \left\{ \mathbf{V} + \lambda_1 \, \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \right\}^{-1} \Gamma^\top \mathbf{S}(\theta_*) \, .$$

Thus, the supremum norm test statistic can be expressed as

$$\omega_{\infty,n} = n^{-1}\lambda_2^{-1}\mathbf{S}(\theta_*)^\top \Gamma \left\{ \mathbf{V} + \lambda_1 \, \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \right\}^{-1} \Gamma^\top \mathbf{S}(\theta_*) \, . \tag{4.16}$$

If we fix λ_1 and λ_2, we obtain a test statistic that can be expressed as a quadratic form in $\mathbf{S}(\theta_*)$. Fixing λ_1 and λ_2 is appealing due to computational difficulties with confidence band construction arising when the test statistic is not available in closed form, as discussed in Section 4.5.4.

One can see in (4.16) that λ_2 affects the scale of $\omega_{\infty,n}$ but has no other effect on its sampling distribution. Thus, there is only a single relevant tuning parameter λ_1, which governs the trade-off between the RKHS norm $J(h)$ and the variance of the Gâteaux derivative estimator $V_0^{-2}(h)$. In fact, with λ_1 fixed, the test statistic is proportional to

$$\sup_{\mathbf{a}} \left\{ \frac{n^{-1}[\mathbf{S}(\theta_*)^\top (\Gamma \mathbf{a})]^2}{\mathbf{a}^\top \mathbf{V} \mathbf{a} + \lambda_1 \mathbf{a}^\top \mathrm{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a}} \right\}$$

and can thus be viewed as a penalized version of the original supremum norm test statistic in (4.13). In our implementation, we use this penalized supremum norm test statistic and take λ_1 as the solution to

$$\frac{\tilde{\mathbf{a}}_{\lambda_1,\lambda_2}^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \tilde{\mathbf{a}}_{\lambda_1,\lambda_2}}{\tilde{\mathbf{a}}_{\lambda_1,\lambda_2}^\top \mathbf{V} \tilde{\mathbf{a}}_{\lambda_1,\lambda_2}} = \gamma.$$

Though this choice is data-adaptive, it is theoretically justified as long as this data-adaptive choice of λ_1 converges in probability to a constant.

To approximate the L_2 norm test statistic $\Omega_2(n^{1/2}\dot{R}_{n,\theta_*})$, we use a Monte Carlo sampling algorithm. Let A be a $N(0, \{\mathbf{V} + \lambda_3 \,\text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right)\}^{-1})$ random variable, where $\lambda_3 > 0$ is a tuning parameter that modulates the distribution of $\frac{A^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) A}{A^\top \mathbf{V} A}$, and let $\mathbf{a}_1, \ldots, \mathbf{a}_B$ be a sample of independent draws from the distribution of A. For large λ_3, $\frac{\mathbf{a}_b^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a}_b}{\mathbf{a}_b^\top \mathbf{V} \mathbf{a}_b}$ will be small for a large proportion of the sample. We select λ_3 so that a large portion of the vectors $\mathbf{a}_b$, say half, belongs to the set $\mathcal{A}_\gamma = \left\{\mathbf{a} : \mathbf{a}^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a} \leq \gamma \mathbf{a}^\top \mathbf{V} \mathbf{a}\right\}$. Let π be an approximation of the density function of $\frac{A^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) A}{A^\top \mathbf{V} A}$, and let $\pi_b = \pi\left(\frac{\mathbf{a}_b^\top \text{diag}\left(\frac{1}{\boldsymbol{\kappa}}\right) \mathbf{a}_b}{\mathbf{a}_b^\top \mathbf{V} \mathbf{a}_b}\right)$. We consider the approximated L_2 norm test statistic

$$\omega_{2,n,B} := \sum_{b=1}^{B} \frac{[\mathbf{S}^\top(\theta_*)\Gamma \mathbf{a}_b]^2}{\mathbf{a}_b^\top \mathbf{V} \mathbf{a}_b} \frac{\mathbb{1}\left(\mathbf{a}_b \in \mathcal{A}_\gamma\right)}{\pi_b} \,. \tag{4.17}$$

The L_2 norm $\omega_{2,n,B}$ is an average of squared Gâteaux derivative estimates for a random sample of directions $h \in \mathcal{H}_\gamma$, inversely weighted by the variance of the estimates and the density of $J(h)V^2(h)$. Inversely weighting by the density of $J(h)V^2(h)$ results in directions with equal smoothness receiving equal weight and reduces the impact of λ_3 on the sampling distribution of the test statistic. The approximated L_2 norm test statistic is also available in quadratic form in $\mathbf{S}(\theta_*)$ as $\omega_{2,n,B} = n^{-1}\left(\Gamma^\top \mathbf{S}(\theta_*)\right)^\top \mathbf{A}^\top \mathbf{U}^{-1} \mathbf{A}\left(\Gamma^\top \mathbf{S}(\theta_*)\right)$, where $\mathbf{A}$ is a $B \times d$ matrix with $(b,j)^{th}$ element $\mathbf{a}_{b,j}$ and $\mathbf{U}$ is a B-dimensional diagonal matrix with b^{th} diagonal entry $\mathbf{U}_b := \pi_b \mathbf{a}_b^\top \mathbf{V} \mathbf{a}_b$.

The distribution of $\mathbf{a}_b$ can influence the statistical power of the test, and alternative approaches for generating the Monte Carlo sample can be considered. For example, if we have prior knowledge about which directions provide strong evidence in favor of the alternative

hypothesis, we may choose to generate Monte Carlo samples from a distribution that places more weight on such directions. When relevant prior knowledge is not available, one may prefer to use the supremum norm.

4.5.3 Calculation of the multiplier bootstrap test statistics

The bootstrap test statistic (4.8) can be computed using a similar strategy as in Section 4.5.2. Let $\xi_1, \xi_2, \ldots, \xi_n$ be a sample of independent draws from the standard normal distribution. The multiplier bootstrap derivative estimate is

$$\dot{R}_{m,n,\theta_n}(h) = \frac{1}{n}\sum_{i=1}^{n} \xi_i \left\{ [Y_i - \mu_{n,Y,P_0}(W_i) + \mu_{n,\theta_n,P_0}(W_i) - \theta_n(X_i)][h(X_i) - \mu_{n,h,P_0}(W_i)] - \dot{R}_{n,\theta_n}(h) \right\}.$$

For $h \in \mathcal{H}_\gamma$, we can re-write $\dot{R}_{m,n,\theta_n}(h)$ as $\mathbf{S}(\theta_n)^\top \text{diag}\left(\boldsymbol{\xi} - \bar{\xi}\right) \Gamma \mathbf{a}$, where $\boldsymbol{\xi} := (\xi_1, \xi_2, \ldots, \xi_n)$, $\bar{\xi} := \frac{1}{n}\sum_{i=1}^{n} \xi_i$, and $\mathbf{S}$, Γ and $\mathbf{a}$ are as defined in Section 4.5.2. Thus, the bootstrap test statistics can be computed using the same routines as discussed in Section 4.5.2, but replacing $\mathbf{S}(\theta_*)$ with $\text{diag}\left(\boldsymbol{\xi} - \bar{\xi}\right) \mathbf{S}(\theta_n)$.

4.5.4 Confidence band construction

As discussed in Section 4.3, to construct confidence bands, the class Θ must be assumed to have sufficient structure. We first discuss the construction of a data-driven approximation to Θ. Similarly as with our construction of the class of directions $\mathcal{H}$, we define our function class using a basis expansion with an additional smoothness constraint,

$$\Theta_\zeta := \left\{ \theta = \textstyle\sum_{j=1}^{d} a_j \eta_j : \sum_{j=1}^{d} \frac{a_j^2}{\kappa_j} < \zeta \right\},$$

where $\zeta > 0$ is a bound on the allowable roughness. As noted in Section 4.3, we require that the smoothness parameter ζ be larger than $J(\theta_0)$ to guarantee that the nominal coverage rate is achieved asymptotically; in practice, we set $\zeta = \zeta_n = J(\theta_n)$. Coverage can be compromised since $J(\theta_n)$ may be a biased estimator of $J(\theta_0)$ when the tuning parameter λ for θ_n in (4.12) is selected so that θ_n is optimal with respect to the mean squared error — van de Geer (2000)

provides a comprehensive discussion of key issues in penalized least squares regression. In simulations, we will see that in the oracle setting, where $\zeta = J(\theta_0)$, nominal coverage is achieved. We also examine the effect of data-adaptive selection on coverage.

We now discuss the computation of confidence bands. For a norm Ω, let t_* be the $(1-\alpha)$-quantile of the limiting distribution of $\Omega(n^{1/2}\dot{R}_{n,\theta_*})$, which can be approximated via the multiplier bootstrap. The upper limit of a confidence interval for the evaluation $\theta_0(x_0)$ of θ_0 at a fixed point x_0 can be taken to be

$$\sup_{a_1,a_2,\ldots,a_d} \left\{ \textstyle\sum_{j=1} a_j \eta_j(x_0) : \sum_{j=1}^d \frac{a_j^2}{\kappa_j} < \zeta,\ \Omega\left(n^{1/2}\dot{R}_{n,\sum_{j=1}^d a_j\eta_j}\right) < t_* \right\}, \tag{4.18}$$

that is, the largest value of $\theta(x_0)$ for $\theta \in \Theta_\zeta$ such that the test statistic is not sufficiently large to reject the hypothesis $\theta_0 = \theta$ against its complement $\theta_0 \neq \theta$. The lower confidence limit takes a similar form but replacing supremum with infimum.

The optimization problem can be challenging if the test statistic $\Omega(n^{1/2}\dot{R}_{n,\sum_j a_j\eta_j})$ is not available as a closed-form function of coefficients $(a_1, a_2, \ldots, a_d)$. We consider the penalized version of the supremum norm test statistic in (4.16) (with λ_1 and λ_2 fixed) and the L_2 norm test statistic in (4.17); both are available as a quadratic form and can be written as $\mathbf{S}\big(\sum_{j=1}^d a_j\eta_j\big)^\top \Pi \mathbf{S}\big(\sum_{j=1}^d a_j\eta_j\big)$, where Π is a matrix that does not depend on the coefficients $a_1, a_2, \ldots, a_d$. For $\mathbf{S}\big(\sum_{j=1}^d a_j\eta_j\big)$ to be available as a closed-form function of $a_1, a_2, \ldots, a_d$, an estimate of the conditional mean of $\theta(X) = \sum_{j=1}^d a_j\eta_j(X)$ given W must also be available in closed form. Similarly as in our construction of estimators of μ_{h,P_0} described in Section 4.5.1, we use $\mu_{n,\theta,P_0} := \sum_{j=1}^d a_j \mu_{n,\eta_j,P_0}$. With this construction, $\mathbf{S}$ is linear in $a_1, a_2, \ldots, a_d$, and so, the optimization problem (4.18) is a quadratically constrained quadratic program that can be solved using interior point methods. Many software packages include implementations for this type of problem — see, for example, R's CVXR package (Fu et al., 2017).

The last main challenge concerns selection of the class $\mathcal{H}$ of directions. To construct confidence bands of optimal width, we must have optimal power to reject each false hypothesis $H : \theta_0 = \theta$. The class of directions that provides the optimal test, however, depends on the null hypothesis — when the distance between θ and θ_0 is larger, the class of directions

should also be larger. While the norm Ω in (4.18) should therefore depend on θ, allowing this dependence can create computational difficulties. We instead take the simple approach of using the same class $\mathcal{H}_\gamma$ to perform all tests. We set $\gamma = \gamma_n = J(\theta_n)V_n^2(\theta_n)$ so that the power of the restricted score test is preserved when the null is flat and θ_0 is non-smooth.

4.6 Results from simulation studies

In this section, we examine the behavior of our proposed methodology in a simulation study. The objectives of this simulation study are to demonstrate that our proposed test provides nominal coverage and type I error rate and, and to examine how the selection of the norm Ω influences statistical power and the width of confidence bands.

4.6.1 Example 1: nonparametric mean regression

We first consider the nonparametric regression setting. We generate synthetic data from the model $Y = \theta_0(X) + \epsilon$, where we set the regression function θ_0 to be

$$\theta_0(x) = \sin[\pi x^2 \text{sign}(x)] \tag{4.19}$$

and we draw independently X from a uniform distribution on $(-1, 1)$ and ϵ from a $N(0, 9)$ distribution. Under these settings, we generate 800 synthetic data sets for $n \in \{100, 200, 300, 400, 500, 1000, 2000\}$. An example data set, with reduced noise, is shown in Figure (4.1).

We compare the restricted score test using the L_2 norm and the quadratic form approximation of the supremum norm described in Section 4.5.2. For confidence band construction, we also consider both known and estimated smoothness of θ_0. We construct $\mathcal{H}$ from a Sobolev basis with $d = 50$ basis functions. For each application of the multiplier bootstrap, we generate 1000 bootstrap samples.

To assess the statistical power, we test the null hypothesis $H_0 : \theta_0 \equiv 0$, and to assess type I error rate control, we test the null at the true value of θ_0. In both cases, we use significance level $\alpha = 0.05$. To assess coverage, we select 100 evenly spaced points on the interval $[-1, 1]$

and determine if the evaluation of the true regression function at each point lies within the confidence band. We summarize the width of the band by calculating the average width at these 50 points.

We compare our restricted score test with the debiased local polynomial regression estimator of Calonico et al. (2018), implemented in the publicly available R package `nprobust` (Calonico et al., 2019). The `nprobust` package is designed for pointwise inference, so we repurposed the method for hypothesis testing and uniform interval construction as we now describe. For a fixed sequence of points $x_1, x_2, \ldots, x_k$, the `nprobust` package outputs an estimate $\check{\boldsymbol{\theta}} := (\check{\theta}(x_1), \check{\theta}(x_2) \ldots, \check{\theta}(x_k))$ of $\boldsymbol{\theta}_0 := (\theta_0(x_1), \theta_0(x_2), \ldots, \theta_0(x_k))$ that approximately satisfies $\check{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, C_1 C_2 C_1)$, where $C_1 := \operatorname{diag}\{\operatorname{sd}(\check{\theta}(x_1)), \operatorname{sd}(\check{\theta}(x_2)), \ldots, \operatorname{sd}(\check{\theta}(x_k))\}$ and $C_2 := \operatorname{corr}(\check{\theta}(x_1), \check{\theta}(x_3), \ldots, \check{\theta}(x_k))$, for sufficiently large n. We use a confidence band of the form

$$\left(\check{\theta}(x_j) - m_{1-\alpha} \operatorname{sd}(\check{\theta}(x_j)),\ \check{\theta}(x_j) + m_{1-\alpha} \operatorname{sd}(\check{\theta}(x_j))\right), \quad j = 1, 2, \ldots, k,$$

where $m_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution for the maximum absolute value of a Gaussian random vector with mean zero and variance C_2. We test $H_0 : \theta_0 = \theta_*$ by verifying whether θ_* resides within the interior of the confidence band, and reject the null hypothesis whenever

$$\max_{j \in \{1,2,\ldots,k\}} \frac{|\check{\theta}(x_j) - \theta_*(x_j)|}{\operatorname{sd}(\check{\theta}(x_j))} > m_{1-\alpha} \ .$$

We summarize results in Figures 4.2, 4.3 and 4.4. The restricted score test provides nominal type I error control in moderate sample sizes when using the L_2 norm and is slightly anti-conservative when using the penalized supremum norm. The restricted score test is also well-powered against the null hypothesis $H_0 : \theta_0 \equiv 0$; the power is comparable for the L_2 and penalized supremum norms, and both choices offer a modest improvement over the debiasing approach. The confidence bands constructed via the restricted score test exceed the nominal coverage rate when the oracle choice smoothness $\zeta = J(\theta_0)$ is supplied. This is unsurprising, as our construction in Section 4.5.2 only guarantees that the coverage rate will be *no smaller than* $(1-\alpha)$. A confidence set for the collection of all functionals of θ_0 should achieve the

nominal coverage rate, but we only consider the subset of evaluation functionals. When using the smoothness of the estimate $\zeta = \zeta_n = J(\theta_n)$, the coverage rate remains close to the nominal level when the supremum norm is used, and exceeds the nominal rate when the L_2 norm is used. The debiasing approach provides the narrowest confidence bands, followed closely by the penalized supremum norm with adaptively-selected smoothness ζ_n and the penalized supremum norm with oracle smoothness. The median upper and lower confidence limits for all methods with $n = 2000$ are provided in Figure 4.4. We find that the confidence bands are able to best capture the shape of the true risk minimizer θ_0 when the supremum norm is used.

We suspect that the width of the confidence bands and their ability to capture the shape of θ_0 depends on the power of the restricted score test to reject null hypotheses in a neighborhood of θ_0. If for any θ sufficiently close to θ_0 there are few directions h such that the Gâteaux derivative evaluated at θ, $\dot{R}_{0,\theta}(h)$, is large, one might expect the restricted score test to have more power to reject the hypothesis $H_0 : \theta_0 = \theta$ when the supremum norm is used instead of the L_2 norm, as discussed in Section 4.4.2. This may explain why the confidence bands obtained using the restricted score test seem to perform best when the supremum norm is used.

4.6.2 Example 2: partially additive mean regression

Our simulation design for the partially additive mean model is similar to the design used in the nonparametric regression setting. We let $W := (W_1, W_2)$ be a vector of two independent uniform random variables on $(-1, 1)$ and subsequently generate $X = \frac{1}{3}W_1 + \frac{1}{3}\sin(\pi W_2) + \Delta$, where Δ follows a uniform distribution on $(-\frac{1}{3}, \frac{1}{3})$. By construction, X has support $(-1, 1)$. We then generate $Y = f_0(W) + \theta_0(X) + \epsilon$, where the nuisance function f_0 is defined pointwise as

$$f_0(w_1, w_2) = -4\left[\frac{\exp(5w_1)}{1 + \exp(5w_1)} - \frac{1}{2}\right] - 2\,\mathrm{sign}(w_2)w_2^2 \ ,$$

the parameter θ_0 of interest is defined as in (4.19), and ϵ is a $N(0, 9)$ random variable independent of (W, X). In Figure (4.5), we show scatter plots of Y against X and $Y - f_0(W)$ against X for an example data set with reduced noise.

We compare our restricted score test under the same settings as in the nonparametric regression example, with the exception that we construct $\mathcal{H}$ using a smaller set of $d = 10$ basis functions. We use the highly adaptive lasso (Benkeser and van der Laan, 2016) to estimate the conditional mean functions μ_{Y,P_0}, μ_{θ_*,P_0} and μ_{h,P_0}. To the best of our knowledge, construction of uniform confidence bands in partially additive models has not been studied, so we do not compare with a competing method.

Simulation results are presented in Figures 4.6, 4.7 and 4.8. Similarly as in the nonparametric regression setting, we find that the restricted score test achieves nominal control of the type I error rate in large sample sizes, and the statistical power is comparable for the both L_2 and penalized supremum norms. For all methods considered, the confidence bands exceed the nominal coverage rate, suggesting that estimating the smoothness of θ_0 does not seriously compromise coverage guarantees. The average width of the confidence bands obtained using the penalized supremum norm and the L_2 norm are similar, and the bands are generally able to capture the shape of θ_0. The width is larger in the tails because X is not uniformly distributed — its distribution is more tightly concentrated around zero.

4.7 Results from the 1987 National Medical Expenditure Survey

We apply our method to data from the 1987 National Medical Expenditure Survey, extracted by Johnson et al. (2003). These data include information about smoking behaviors and medical expenditures in a sample of 9,708 US citizens. The objective of this analysis is to assess the association between tobacco exposure, measured in pack-years (the number of cigarette packs an individual smoked per day times the number of years they smoked), and medical expenditure.

We fit a partially additive mean model in which where the outcome Y is the total medical expenditure, the exposure X is the natural log of the number of pack-years, and the

adjustment covariate vector W includes: participant age at the time of the survey, age at initiation of smoking, gender, marital status, education level, census region, and socioeconomic status. We are interested in testing the null hypothesis that there is no association between pack-years smoked and medical expenditures ($\theta_0 \equiv 0$) and in constructing a confidence band for θ_0. We apply the restricted score test using the approximate supremum norm for testing and confidence band construction. As in the simulation study, we construct the class of directions $\mathcal{H}$ using a Sobolev basis with $d = 10$ basis functions.

Results from the analysis are summarized in Figures 4.9 and 4.10. Figure 4.9 shows a scatter plot of log-pack-years against medical expenditure, and suggests that the two variables share a positive marginal association. In Figure 4.10, we present our estimate of the mean regression function from the partially additive model and a 95% confidence band. We find a strongly significant association between pack-years and medical expenditure, with p-value $p = 0.001$ based on 10,000 bootstrap samples. The resulting estimate of the regression function suggests that a small amount of smoking (less than 2 log-pack-years, i.e., 7.5 pack-years) is not strongly associated with an increased average medical expenditure, though there is an increase in average medical expenditure associated with a moderate or large amount of exposure to smoking.

4.8 Discussion

We have introduced a general approach for hypothesis testing and constructing confidence bands for infinite-dimensional parameters in nonparametric and semiparametric statistical models. Our framework is applicable to any function-valued parameter that can be expressed as a risk minimizer. While we consider the nonparametric and semiparametric partially additive mean regression models as examples in this chapter, the framework can be useful in other widely relevant applications. For example, in causal inference, both the conditional average treatment effect curve and the causal dose-response function are parameters tat can be expressed expressed as risk minimizers, so that inference can be conducted using the restricted score test.

There remain several important issues that must be clarified through additional research. In order to guarantee adequate coverage for resulting confidence bands, our approach requires an upper bound for the smoothness of the true risk minimizer, and this requirement can be seen as a limitation of the method. However, we found in simulations that using a naive plug-in estimator may work well in practice, even if such estimators are not necessarily consistent. Regardless, our confidence band with estimated smoothness retains a meaningful interpretation as an envelope containing a set of smooth functions consistent with the observed data. Additional work is also required to provide further guidance on the selection of the class of directions $\mathcal{H}$ and the choice of norm Ω. Though these choices do not affect type I error rate, they can influence statistical power and the width of resulting confidence bands. For the purposes of this chapter, we have only provided some heuristic guidance for selecting a class of directions, and leave more extensive and rigorous studies of this question to future work.

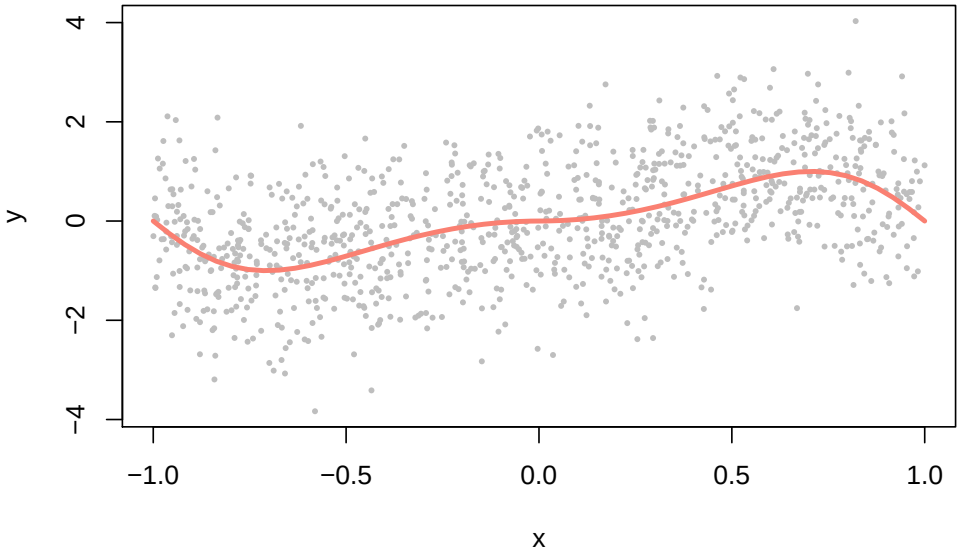


Figure 4.1: Scatter plot of example data set generated under simulation settings described in Section 4.6.1, but with a reduction in noise. The pink curve represents θ_0.

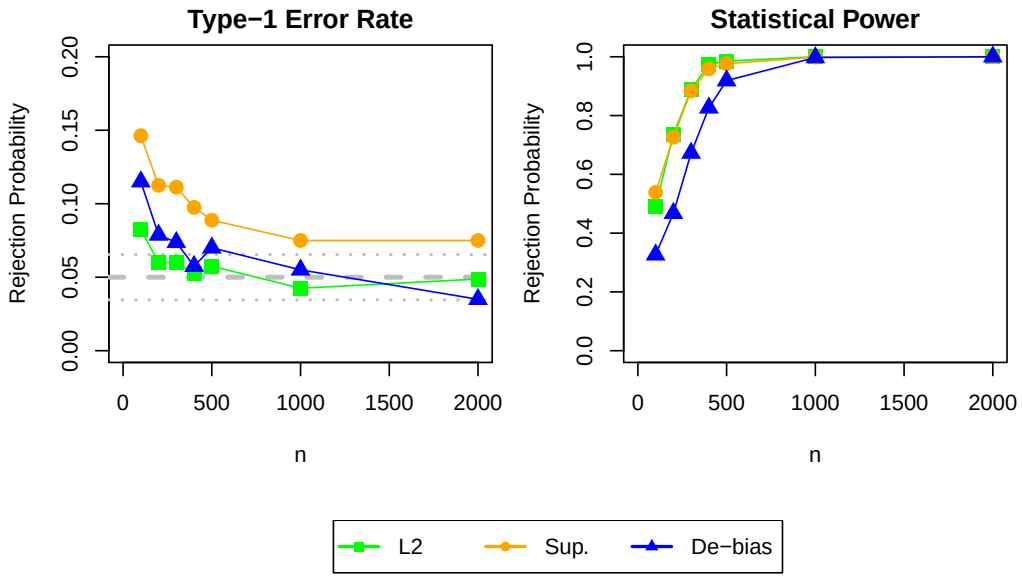


Figure 4.2: Monte Carlo estimates of type-1 error rate (left) and statistical power (right) in the regression setting with the significance level $\alpha = .05$. The dashed gray line indicates the significance level, and the dotted gray lines are placed two Monte Carlo standard errors above and below.

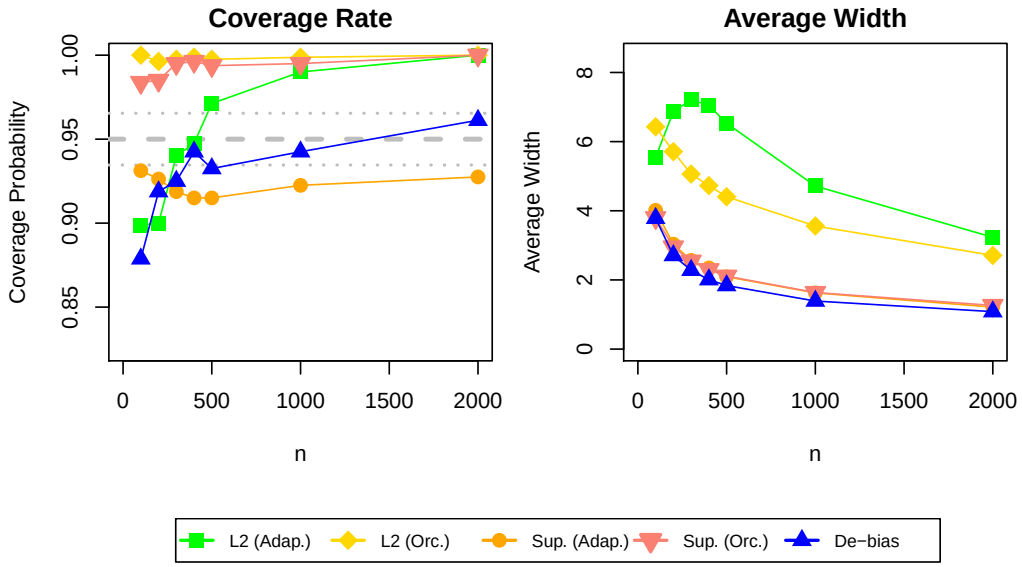


Figure 4.3: Monte Carlo estimates of the coverage probability (left) and average width of confidence band (right) in the regression setting. The dashed gray line indicates the nominal coverage rate .95, and the dotted gray lines are placed two Monte Carlo standard errors above and below.

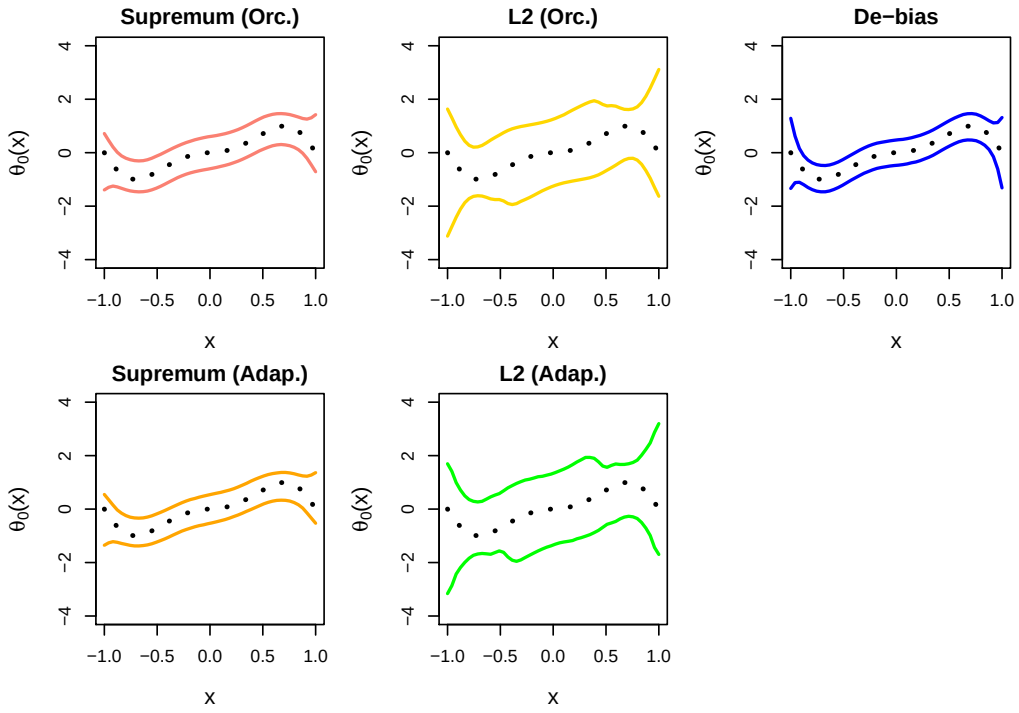


Figure 4.4: Median upper and lower limits of confidence bands with $n = 2000$ in the regression setting. The dotted black line represents the true risk minimizer θ_0.

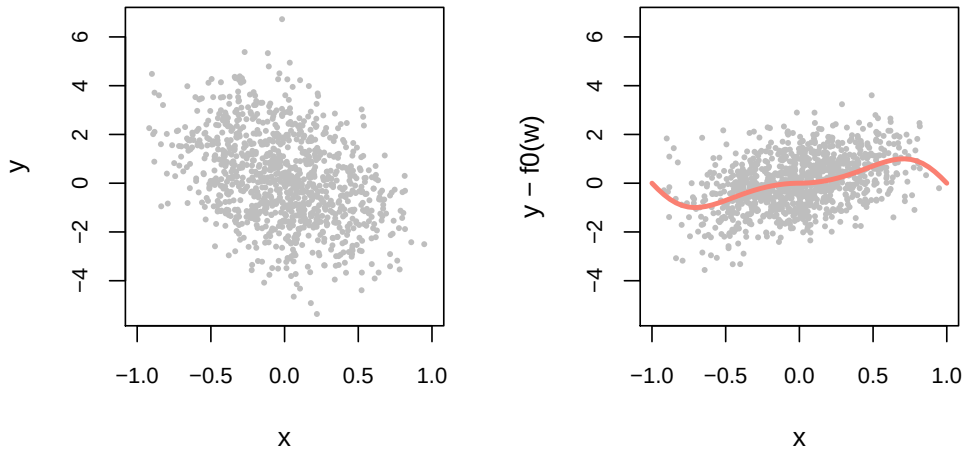


Figure 4.5: Scatter plots of example data set generated under simulation settings described in Section 4.6.2, but with a reduction in noise. The pink curve represents θ_0.

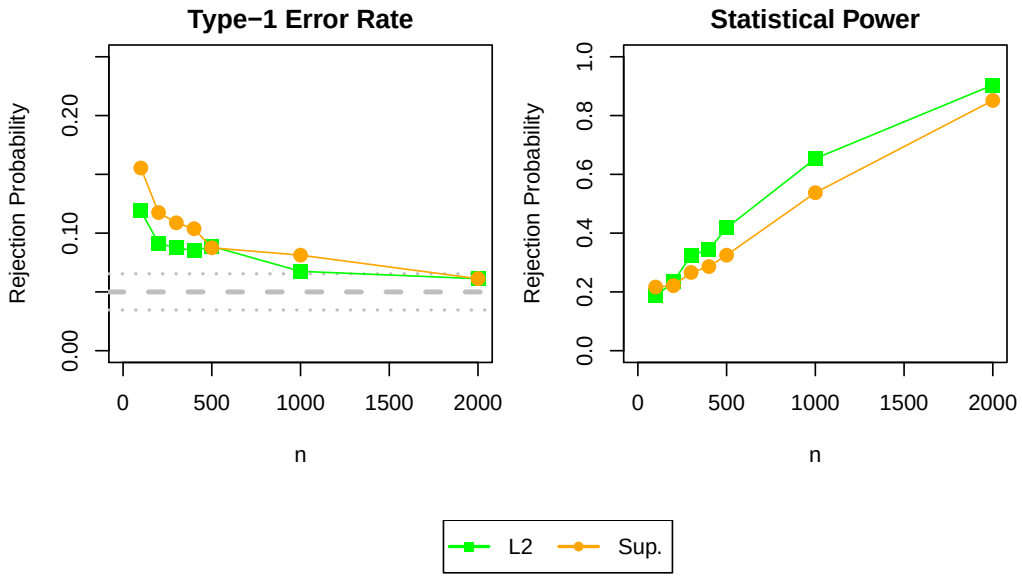


Figure 4.6: Monte Carlo estimates of type-1 error rate (left) and statistical power (right) for the partially additive model with the significance level $\alpha = .05$. The dashed gray line indicates the significance level, and the dotted gray lines are placed two Monte Carlo standard errors above and below.

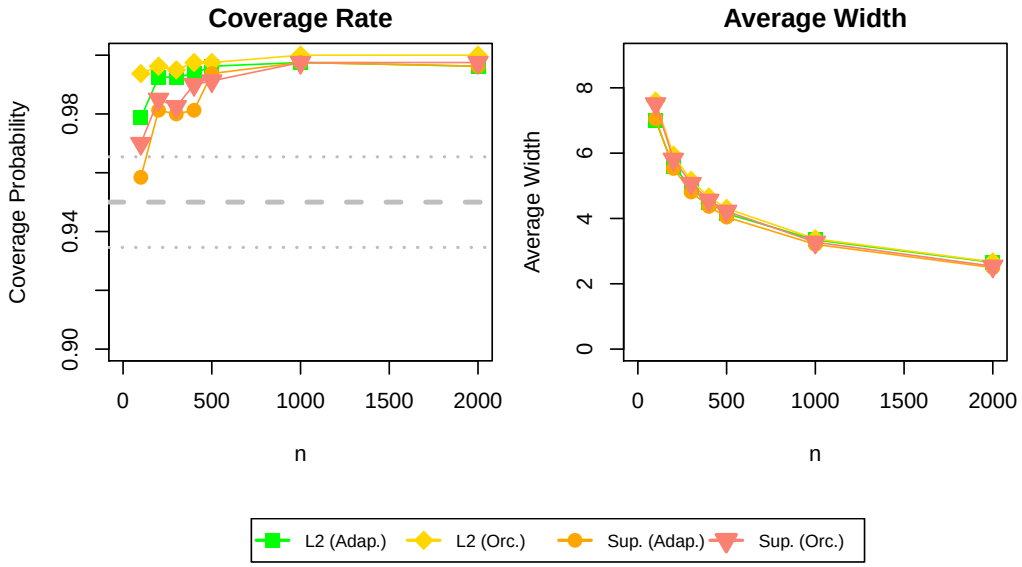


Figure 4.7: Monte Carlo estimates of the coverage probability (left) and average width of confidence band (right) for the partially additive model. The dashed gray line indicates the nominal coverage rate .95, and the dotted gray lines are placed two Monte Carlo standard errors above and below.

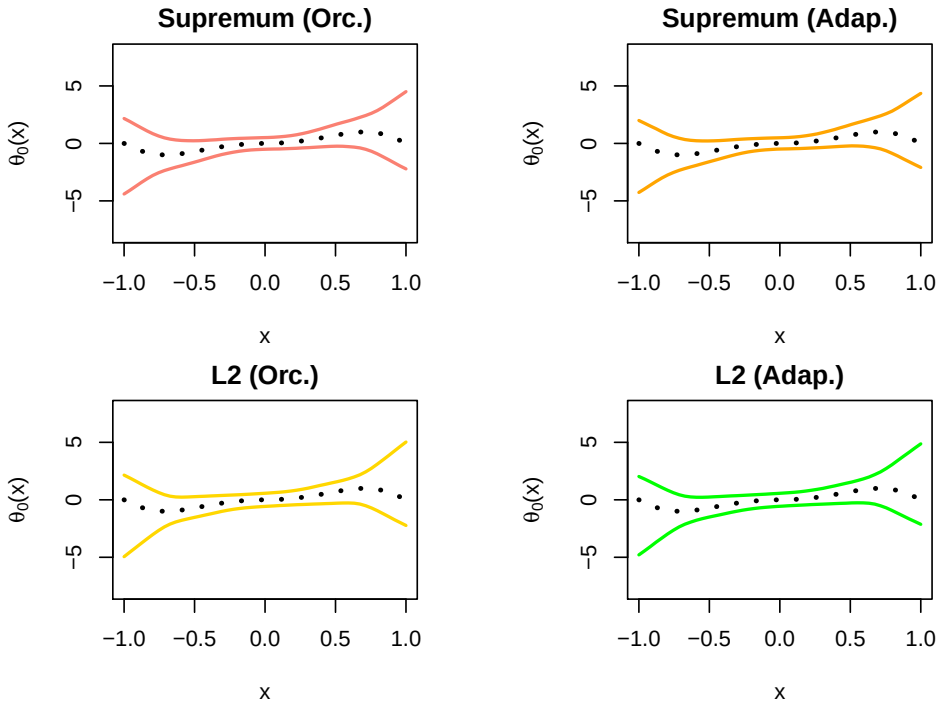


Figure 4.8: Median upper and lower limits of confidence bands with $n = 2000$ for the partially additive model.

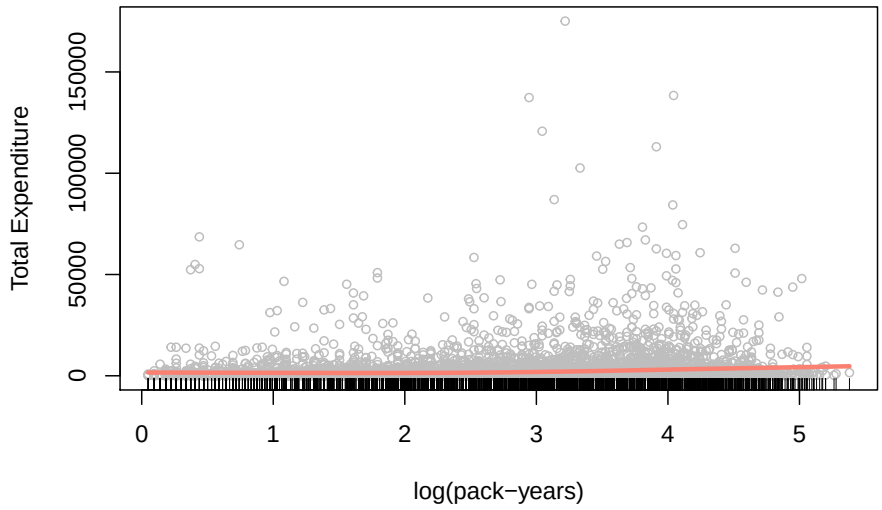


Figure 4.9: Scatter plot of outcome (total medical expenditure) and main exposure (log-pack years) in the 1987 National Medical Expenditure Survey data. The pink curve represents a smoothing spline estimate of the regression function in an unadjusted model.

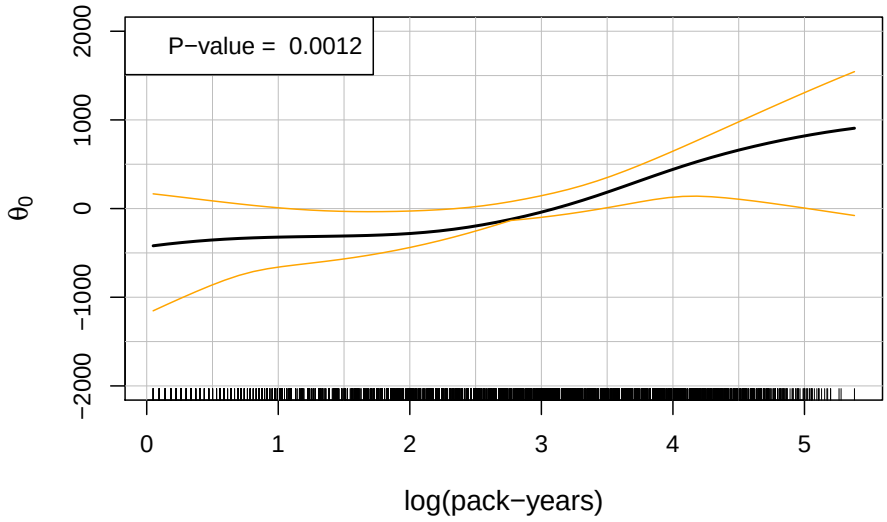


Figure 4.10: Estimate and confidence band for partially additive model fit to the 1987 National Medical Expenditure Survey data.

Chapter 5

DISCUSSION

5.1 Summary

In this dissertation, we proposed solutions to three statistical inference problems. We first considered two challenges that arise when studying heterogeneous effects. In Chapter 2, we discussed approaches for evaluating the hypothesis that an effect is zero in one sub-population and non-zero in another. We argued that this hypothesis is only testable under untenable assumptions and proposed a method for approximating the problem by making inference on the relative difference in effect size. In Chapter 3, we illustrated how differential network analysis can be impacted when the inter-node associations depend on covariates. We showed that failing to adjust for covariates can in some settings result in an inflated type-1 rate or reduced power to detect differential connections. We also developed a covariate-adjusted hypothesis test for differential network connectivity that can be applied to a broad class of parametric probabilistic graphical models in the low- and high-dimensional settings.

We then in Chapter 4 proposed an approach to inference on complex infinite-dimensional estimands in nonparametric and semiparametric models. When the target estimand is non-pathwise differentiable, i.e., it cannot be characterized as a smooth functional of the underlying probability distribution, it is challenging to conduct formal inference because efficient estimators generally do not attain tractable limiting distributions. We developed an inferential procedure that does not require characterizing the sampling distribution of the estimator, thereby circumventing this problem. In many instances, the target estimand can be characterized as the minimizer of a population risk functional. This representation can be used to derive a set of estimating equations that any minimizer of the risk must satisfy. One can therefore test the null hypothesis that a candidate function is the true risk minimizer by

verifying whether the estimating equations are satisfied by the null parameter. Furthermore, one can construct a confidence band for the risk minimizer by characterizing the set of functions that the hypothesis test does not reject. Our novel strategy has wide application in a variety of inference problems and requires very few assumptions about the data-generating mechanism.

5.2 Future Work

Several potential research directions can be explored based on the work presented in this dissertation. The inferential framework we present in Chapter 2 is only suitable when the objective is to study effect heterogeneity with two sub-populations. It may be of interest to extend this framework to be applicable when there are more than two groups, or when sub-populations are defined by a continuous variable. In Chapter 3, we assume a parametric form for the probabilistic graphical models, and there is potential for model misspecification as such parametric assumptions often fail in practice. I am interested in relaxing these assumptions and considering a richer class of semiparametric or nonparametric graphical models. We also assume that the dependency of the inter-node associations on the covariates can be modeled parametrically. We could alternatively use a more flexible model and explore how one could perform nonparametric inference, perhaps using the restricted score test described in Chapter 4.

There several avenues of research stemming from our work in Chapter 4 that I am interested in pursuing. While our inferential framework is widely-applicable, we only considered some pedagogical examples in Chapter 4 and have not yet fully demonstrated the framework's utility. I would like to implement the restricted score test to treat additional examples, such as the conditional average treatment effect and the causal dose-resposne function, that are of greater practical interest. In addition, I plan to develop an open-access software package to make our proposed methodology accessible to a wide audience. I also would like to extend the framework to address more challenging methodologic problems such as testing with composite null hypotheses and inference for risk minimizers when the risk functional is not

differentiable.

There are also several theoretical questions about the restricted score test that are worth exploring. I would like to analytically characterize the power of the restricted score test so that it can be better understood how tuning parameter selection influences statistical power. It would also be interesting to determine if there are settings in which the restricted score test is the most powerful test and if the confidence bands obtained via the restricted score test are rate optimal. Additionally, I would like to investigate how the restricted score test relates to more classical parametric inference. In parametric models, there are three commonly-used approaches for likelihood-based inference: the Wald test, the likelihood ratio test, and the score test. All three tests have been shown to be asymptotically equivalent. As the restricted score test is a generalization of the parametric score test, it would be unsurprising if the restricted score test could be related to generalized versions of the Wald and likelihood ratio tests. I am interested in further exploring these connections to gain insight into the behavior of the restricted score test.

BIBLIOGRAPHY

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12,** 56–68.

Barron, A. R. (1989). Statistical properties of artificial neural networks. In *Proceedings of the 28th IEEE Conference on Decision and Control,*, pages 280–285. IEEE.

Bayarri, M. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association* **95,** 1127–1142.

Belilovsky, E., Varoquaux, G., and Blaschko, M. B. (2016). Testing for differences in Gaussian graphical models: Applications to brain connectivity. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* pages 1165–1188.

Benkeser, D. and van der Laan, M. (2016). The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its Interface* **2,** 369.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B* pages 349–372.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association* **113,** 767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv preprint arXiv:1906.00198* .

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. A., Tse, C. K., Edmiston, S., et al. (2006). Race, breast cancer subtypes, and survival in the carolina breast cancer study. *Journal of the American Medical Association* **295,** 2492–2502.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102,** 47–64.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21,** C1–C68.

Daly, R. J., Binder, M. D., and Sutherland, R. L. (1994). Overexpression of the Grb2 gene in human breast cancer cell lines. *Oncogene* **9,** 2723–2727.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B* **76,** 373–397.

de Gonzalez, A. B., Cox, D. R., et al. (2007). Interpretation of interaction: A review. *Annals of Applied Statistics* **1,** 371–385.

de la Fuente, A. (2010). From 'differential expression'to 'differential networking'– identification of dysfunctional regulatory networks in diseases. *Trends in Genetics* **26,** 326–333.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* **88,** 1287–1301.

Farabaugh, S. M., Boone, D. N., and Lee, A. V. (2015). Role of IGF1R in breast cancer subtypes, stemness, and lineage differentiation. *Frontiers in endocrinology* **6,** 59.

Fieller, E. C. (1940). The biological standardization of insulin. *Supplement to the Journal of the Royal Statistical Society* **7,** 1–64.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441.

Fu, A., Narasimhan, B., and Boyd, S. (2017). CVXR: An R package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582* .

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41,** 361–372.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98,** 1–15.

Ha, M. J., Banerjee, S., Akbani, R., Liang, H., Mills, G. B., Do, K.-A., and Baladandayuthapani, V. (2018). Personalized integrated network modeling of the cancer proteome atlas. *Scientific reports* **8,** 1–14.

Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22,** 215–232.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics* **20,** 675–694.

Hall, P., Horowitz, J., et al. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics* **41,** 1892–1921.

Hardle, W. and Marron, J. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics* **19,** 778–796.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B* **55,** 757–779.

He, H., Cao, S., Zhang, J.-g., Shen, H., Wang, Y.-P., and Deng, H. (2019). A statistical test for differential network analysis based on inference of Gaussian graphical model. *Scientific Reports* **9,** 1–8.

Homrighausen, D. and McDonald, D. J. (2017). Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica* pages 1017–1036.

Honda, T. (2019). The de-biased group lasso estimation for varying coefficient models. *Annals of the Institute of Statistical Mathematics* **73,** 1–27.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* **6,** 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics & Data Analysis* **51,** 2499–2512.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology* **8,** 565.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15,** 2869–2909.

Jin, Z., Tamura, G., Tsuchiya, T., Sakata, K., Kashiwaba, M., Osakabe, M., and Motoyama, T. (2001). Adenomatous polyposis coli (apc) gene promoter hypermethylation in primary breast cancers. *British Journal of Cancer* **85,** 69–73.

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112,** 135–151.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28,** 27–30.

Khan, S. A., Rogers, M. A., Khurana, K. K., Meguid, M. M., and Numann, P. J. (1998). Estrogen receptor expression in benign breast epithelium and breast cancer risk. *Journal of the National Cancer Institute* **90,** 37–42.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Kurozumi, S., Matsumoto, H., Hayashi, Y., Tozuka, K., Inoue, K., Horiguchi, J., Takeyoshi, I., Oyama, T., and Kurosumi, M. (2017). Power of PgR expression as a prognostic factor for er-positive/her2-negative breast cancer patients at intermediate risk classified by the ki67 labeling index. *BMC Cancer* **17,** 354.

Li, J. and Chan, I. S. (2006). Detecting qualitative interactions in clinical trials: an extension of range test. *Journal of Biopharmaceutical Statistics* **16,** 831–841.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics* **10,** 806–854.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10,** 2295–2328.

Lu, J., Kolar, M., and Liu, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association* **115,** 2084–2099.

Lumachi, F., Brunello, A., Maruzzo, M., Basso, U., and Mm Basso, S. (2013). Treatment of estrogen receptor-positive breast cancer. *Current Medicinal Chemistry* **20,** 596–604.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

McNamee, R. (2005). Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* **62,** 500–506.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34,** 1436–1462.

Mitra, R. and Zhang, C.-H. (2016). The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics* **10,** 1829–1873.

Müller, A. (2001). Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics* **53,** 567–575.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27,** 538–557.

Neumann, M. H. et al. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *The Annals of Statistics* **23,** 1937–1959.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review* **45,** 167–256.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019). Bayesian graphical regression. *Journal of the American Statistical Association* **114,** 184–197.

Pan, G. and Wolfe, D. A. (1997). Test for qualitative interaction of clinical significance. *Statistics in Medicine* **16,** 1645–1652.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*. Springer.

Piantadosi, S. and Gail, M. (1993). A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* **12,** 1239–1248.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press.

Reinert, T., Coelho, G. P., Mandelli, J., Zimermann, E., Zaffaroni, F., Bines, J., Barrios, C. H., and Graudenz, M. S. (2019). Association of esr1 mutations and visceral metastasis in patients with estrogen receptor-positive advanced breast cancer from brazil. *Journal of oncology* **2019,**.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* **56,** 931–954.

Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics* **10,** 1341–1392.

Shojaie, A. (2020). Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics* page e1508.

Silvapulle, M. J. (2001). Tests against qualitative interaction: Exact critical values and robust tests. *Biometrics* **57,** 1157–1165.

Sun, J., Loader, C. R., et al. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics* **22,** 1328–1345.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58,** 267–288.

van de Geer, S. (2000). *Empirical processes in M-estimation*, volume 6. Cambridge university press.

van de Geer, S. (2016). Estimation and testing under sparsity. *Lecture Notes in Mathematics* **2159,**.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42,** 1166–1202.

van der Laan, M. J., Bibaut, A., and Luedtke, A. R. (2018). CV-TMLE for nonpathwise differentiable target parameters. In *Targeted Learning in Data Science*, pages 455–481. Springer.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology* **6,**.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

van der Laan, M. J. and Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

VanderWeele, T. J. (2019). The interaction continuum. *Epidemiology (Cambridge, Mass.)* **30,** 648.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104,** 747–757.

Wang, J. and Kolar, M. (2014). Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638* .

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45,** 1113–1120.

Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102,** 247–266.

Xia, Y., Cai, T., and Cai, T. T. (2018). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statistica Sinica* **28,** 63–92.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research* **16,** 3813–3847.

Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V., Schadt, E. E., Zhu, J., et al. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific Reports* **5,** 1–16.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing* **25,** 1129–1141.

Yu, M., Gupta, V., and Kolar, M. (2020). Simultaneous inference for pairwise graphical models with generalized score matching. *Journal of Machine Learning Research* **21,** 1–51.

Yu, S., Drton, M., and Shojaie, A. (2019). Generalized score matching for non-negative data. *Journal of Machine Learning Research* **20,** 1–70.

Yu, S., Drton, M., and Shojaie, A. (2021). Generalized score matching for general domains. *Information and Inference: A Journal of the IMA* .

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68,** 49–67.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76,** 217–242.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association* **112,** 757–768.

Zhao, S. D., Cai, T. T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika* **101,** 253–268.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning* **80,** 295–319.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67,** 301–320.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Theoretical Results for Relative Difference Likelihood Ratio Test

Below, we generalize Lemma 2.1 and Proposition 2.1 to the setting of unbalanced sample sizes $n_1 \neq n_2$.

Lemma A.1. *The likelihood ratio test statistic T^κ can be written as*

$$T^\kappa = \frac{\hat{\theta}_{\max} - \kappa\hat{\theta}_{min}}{\sqrt{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}}},$$

where $\hat{\tau}_{\max} = n_1^{-1}\sigma_1^2\mathbb{1}(|\theta_1| = \theta_{\max}) + n_2^{-1}\sigma_2^2\mathbb{1}(|\theta_2| = \theta_{\max})$, and $\hat{\tau}_{\min} = n_1^{-1}\sigma_1^2\mathbb{1}(|\theta_1| = \theta_{\min}) + n_2^{-1}\sigma_2^2\mathbb{1}(|\theta_2| = \theta_{\min})$.

Proof of Lemma A.1. We define $\hat{\theta}_{\text{ord}}$ and $\hat{\Sigma}_{\text{ord}}$ as

$$\hat{\theta}_{\text{ord}} = \begin{pmatrix} \hat{\theta}_{\max} \\ \hat{\theta}_{\min} \end{pmatrix} \qquad \hat{\Sigma}_{\text{ord}} = \begin{pmatrix} \hat{\tau}_{\max} & 0 \\ 0 & \hat{\tau}_{\min} \end{pmatrix}.$$

Now, we define θ^*_{ord} as the projection of $\hat{\theta}_{\text{ord}}$ onto the null region. If $\hat{\theta}$ is in the null region Θ_0^κ, the projection is equal to $\hat{\theta}_{\text{ord}}$. Otherwise, θ^*_{ord} is the projection of $\hat{\theta}_{\text{ord}}$ onto the space spanned by $(\kappa, 1)^\top$, with distance inversely weighted by Σ_{ord}. We proceed by finding an

expression for θ^*_{ord}.

$$\theta^*_{\text{ord}} = \begin{pmatrix}\kappa\\1\end{pmatrix}\left[\begin{pmatrix}\kappa & 1\end{pmatrix}\begin{pmatrix}\hat{\tau}_{\max} & 0\\0 & \hat{\tau}_{\min}\end{pmatrix}\begin{pmatrix}\kappa\\1\end{pmatrix}\right]^{-1}\begin{pmatrix}\kappa & 1\end{pmatrix}\begin{pmatrix}\hat{\tau}_{\max} & 0\\0 & \hat{\tau}_{\min}\end{pmatrix}\begin{pmatrix}\hat{\theta}_{\max}\\\hat{\theta}_{\min}\end{pmatrix} = \frac{\kappa\hat{\tau}_{\max}\hat{\theta}_{\max}+\hat{\tau}_{\min}\hat{\theta}_{\min}}{\kappa^2\hat{\tau}_{\max}+\hat{\tau}_{\min}}\begin{pmatrix}\kappa\\1\end{pmatrix}.$$

Now, the weighted distance between $\hat{\theta}_{\text{ord}}$ and θ^*_{ord} is

$$\begin{aligned}&(\hat{\theta}_{\text{ord}}-\theta^*_{\text{ord}})^\top\Sigma^{-1}_{\text{ord}}(\hat{\theta}_{\text{ord}}-\theta^*_{\text{ord}}) =\\&\hat{\tau}^{-1}_{\max}\left(\hat{\theta}_{\max}-\frac{\kappa^2\hat{\tau}_{\max}\hat{\theta}_{\max}+\kappa\hat{\tau}_{\min}\hat{\theta}_{\min}}{\kappa^2\hat{\tau}_{\max}+\hat{\tau}_{\min}}\right)^2+\hat{\tau}^{-1}_{\min}\left(\hat{\theta}_{\min}-\frac{\kappa\hat{\tau}_{\max}\hat{\theta}_{\max}+\hat{\tau}_{\min}\hat{\theta}_{\min}}{\kappa^2\hat{\tau}_{\max}+\hat{\tau}_{\min}}\right)^2 =\\&\hat{\tau}^{-1}_{\max}\left(\frac{\hat{\tau}^{-1}_{\min}\hat{\theta}_{\max}-\kappa\hat{\tau}^{-1}_{\min}\hat{\theta}_{\min}}{\kappa^2\hat{\tau}^{-1}_{\max}+\hat{\tau}^{-1}_{\min}}\right)^2+\hat{\tau}^{-1}_{\min}\left(\frac{\kappa\hat{\tau}^{-1}_{\max}\hat{\theta}_{\max}+\kappa^2\hat{\tau}^{-1}_{\max}\hat{\theta}_{\min}}{\kappa^2\hat{\tau}^{-1}_{\max}+\hat{\tau}^{-1}_{\min}}\right)^2 =\\&\frac{\kappa^2\hat{\tau}^{-2}_{\max}\hat{\tau}^{-1}_{\min}+\hat{\tau}^{-1}_{\max}\hat{\tau}^{-2}_{\min}}{(\kappa^2\hat{\tau}^{-1}_{\max}+\hat{\tau}^{-1}_{\min})^2}(\hat{\theta}_{\max}-\kappa\hat{\theta}_{\min})^2 =\\&\frac{(\hat{\theta}_{\max}-\kappa\hat{\theta}_{\min})^2}{\hat{\tau}_{\max}+\kappa^2\hat{\tau}_{\min}}.\end{aligned}$$

Hence, the likelihood ratio test rejects the null for large values of

$$\frac{(\hat{\theta}_{\max}-\kappa\hat{\theta}_{\min})^2}{\hat{\tau}_{\max}+\kappa^2\hat{\tau}_{\min}}\mathbb{1}\left(\hat{\theta}_{\max}-\kappa\hat{\theta}_{\min}>0\right),$$

or equivalently, for large positive values of T^κ, where

$$T^\kappa=\frac{\hat{\theta}_{\max}-\kappa\hat{\theta}_{\min}}{\sqrt{\hat{\tau}_{\max}+\kappa^2\hat{\tau}_{\min}}}.$$

□

Proposition A.1. *Let $n_1+n_2=N$, and assume $n_1/N\to\lambda>0$ as $n_1,n_2\to\infty$. Then,*

i In the interior of the null region, i.e., when $\theta_{\max} - \kappa\theta_{\min} < 0$, T^κ *converges in distribution to* $-\infty$.

ii At all nonzero boundary points of the null region, i.e., when $\theta_{\max} = \kappa\theta_{\min} > 0$, T^κ *converges in distribution to a standard normal random variable.*

Proof of Proposition A.1. First we prove (i). Suppose $\theta_{\max} - \kappa\theta_{\min} = b < 0$. Then consistency of $\hat{\theta}$ and the continuous mapping theorem imply that $\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min} \to_p b$. Now, because $n_1^{-1}\sigma_1^2$ and $n_2^{-1}\sigma_2^2$ both tend to zero, $\frac{1}{\sqrt{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}}} \to_p \infty$. Therefore, $T^\kappa \to_p -\infty$.

Now we prove (ii). Suppose $\theta_{\max} = \kappa\theta_{\min} > 0$. By applying the delta method,

$$\sqrt{\frac{n_1 n_2}{N}}\left(\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min}\right) =$$
$$\sqrt{\frac{n_1 n_2}{N}}\left[\left(\max\left\{|\hat{\theta}_1|, |\hat{\theta}_2|\right\} - \kappa\min\left\{|\hat{\theta}_1|, |\hat{\theta}_2|\right\}\right) - \left(\max\{|\theta_1|, |\theta_2|\} - \kappa\min\{|\theta_1|, |\theta_2|\}\right)\right] =$$
$$\sqrt{\frac{n_1 n_2}{N}} M \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{pmatrix} + o_p(1),$$

where the matrix M is

$$M = \begin{pmatrix} (-\kappa)^{\mathbb{1}(|\theta_1| = \theta_{\min})}\operatorname{sign}(\theta_1) & 0 \\ 0 & (-\kappa)^{\mathbb{1}(|\theta_2| = \theta_{\min})}\operatorname{sign}(\theta_2) \end{pmatrix}.$$

Now, Slutsky's theorem implies that

$$\sqrt{\frac{n_1 n_2}{N}} M \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{pmatrix} + o_p(1) \to_d N\left(0, (1-\lambda)\left(\kappa^2\right)^{\mathbb{1}(|\theta_1| = \theta_{\min})}\sigma_1^2 + \lambda\left(\kappa^2\right)^{\mathbb{1}(|\theta_2| = \theta_{\min})}\sigma_2^2\right).$$

The test statistic T^κ can be written as

$$\begin{aligned} T^\kappa &= \frac{\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min}}{\sqrt{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}}} \\ &= \frac{\sqrt{\frac{n_1 n_2}{N}}\left(\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min}\right)}{\sqrt{\frac{n_1 n_2}{N}\hat{\tau}_{\max}^2 + \frac{n_1 n_2}{N}\kappa^2\hat{\tau}_{\min}^2}}, \end{aligned}$$

with $\hat{\tau}_{\max}$ and $\hat{\tau}_{\min}$ as defined in Proposition A.1. By the continuous mapping theorem, the denominator converges in probability to

$$\sqrt{(1-\lambda)\left(\kappa^2\right)^{\mathbb{1}(|\theta_1|=\theta_{\min})}\sigma_1^2 + \lambda\left(\kappa^2\right)^{\mathbb{1}(|\theta_2|=\theta_{\min})}\sigma_2^2}.$$

Thus, Slutsky's theorem implies that

$$T^{\kappa} \to_d N(0,1).$$

□

In Proposition A.2, we characterize the local asymptotic behavior of the relative difference likelihood ratio test.

Proposition A.2. *Let $n_1 + n_2 = N$, and assume $n_1/N \to \lambda > 0$ as $n_1, n_2 \to \infty$. Suppose $\theta_1 = n_1^{-1/2}c_1$, and $\theta_2 = n_2^{-1/2}c_2$. Further, assume that $\hat{\theta}$ is locally regular in the sense that $\sqrt{n_g}\hat{\theta}_g \to_d N(c_g, \sigma_g^2)$ for $g \in \{1,2\}$. Then as $n_1, n_2 \to \infty$, for all $t > 0$, $\mathbb{P}\left(T^{\kappa} > t\right)$ converges to*

$$\mathbb{P}(W_{11} > t, W_{12} > t) + \mathbb{P}(W_{11} < -t, W_{12} < -t) + \mathbb{P}(W_{21} > t, W_{22} > t) + \mathbb{P}(W_{21} < -t, W_{22} < -t),$$

where (W_{11}, W_{12}) follows a bivariate normal distribution with mean $(\tilde{c}_{11}, \tilde{c}_{12})$, unit variance and correlation ν_1, and (W_{21}, W_{22}) follows a bivariate normal distribution with mean $(\tilde{c}_{21}, \tilde{c}_{22})$, unit variance and correlation ν_2, with $\tilde{c}_{11}, \tilde{c}_{12}, \tilde{c}_{21}, \tilde{c}_{22}, \nu_1, \nu_2$ are defined as

$$\begin{aligned}
\tilde{c}_{11} &= \frac{(1-\lambda)c_1 - \kappa\lambda c_2}{\sqrt{(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2}} \\
\tilde{c}_{12} &= \frac{(1-\lambda)c_1 + \kappa\lambda c_2}{\sqrt{(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2}} \\
\tilde{c}_{21} &= \frac{\lambda c_2 - \kappa(1-\lambda)c_1}{\sqrt{\kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}} \\
\tilde{c}_{22} &= \frac{\lambda c_2 + \kappa(1-\lambda)c_1}{\sqrt{\kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}} \\
\nu_1 &= \frac{(1-\lambda)\sigma_1^2 - \kappa^2\lambda\sigma_2^2}{(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2} \\
\nu_2 &= \frac{\lambda\sigma_2^2 - \kappa^2(1-\lambda)\sigma_1^2}{\kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}.
\end{aligned}$$

Remark A.1. *Setting $c_1 = c_2 = 0$, we obtain the limiting distribution of T^κ when $\theta_1 = \theta_2 = 0$, which is critical for calculating the p-value $\rho^\kappa(t)$.*

Proof of Proposition A.2. We first re-write the tail probability as

$$\mathbb{P}\left(\frac{\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min}}{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}} > t\right) =$$

$$\begin{aligned}
&\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} > t, \frac{\hat{\theta}_1 + \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} > t\right) + \\
&\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} < -t, \frac{\hat{\theta}_1 + \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} < -t\right) + \\
&\mathbb{P}\left(\frac{\hat{\theta}_2 - \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} > t, \frac{\hat{\theta}_2 + \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} > t\right) + \\
&\mathbb{P}\left(\frac{\hat{\theta}_2 - \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} < -t, \frac{\hat{\theta}_2 + \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} < -t\right).
\end{aligned}$$

Thus,

$$\sqrt{\frac{n_1 n_2}{N}}\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix} \to_d N\left(\begin{pmatrix}(1-\lambda)c_1\\ \lambda c_2\end{pmatrix}, \begin{pmatrix}(1-\lambda)\sigma_1^2 & 0\\ 0 & \lambda\sigma_2^2\end{pmatrix}\right).$$

And,

$$\sqrt{\frac{n_1 n_2}{N}}\begin{pmatrix}1 & -\kappa\\ 1 & \kappa\end{pmatrix}\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix} \to_d N\left(\begin{pmatrix}(1-\lambda)c_1 - \kappa\lambda c_2\\ (1-\lambda)c_1 + \kappa\lambda c_2\end{pmatrix}, \begin{pmatrix}(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2 & (1-\lambda)\sigma_1^2 - \kappa^2\lambda\sigma_2^2\\ (1-\lambda)\sigma_1^2 - \kappa^2\lambda\sigma_2^2 & \kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2\end{pmatrix}\right).$$

Now,

$$\begin{pmatrix}\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\\ \frac{\hat{\theta}_1 + \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\end{pmatrix} = \begin{pmatrix}\frac{\sqrt{\frac{n_1 n_2}{N}}(\hat{\theta}_1 - \kappa\hat{\theta}_2)}{\sqrt{\kappa^2 n_2 N^{-1}\sigma_1^2 + n_1 N^{-1}\sigma_2^2}}\\ \frac{\sqrt{\frac{n_1 n_2}{N}}(\kappa\hat{\theta}_1 - \hat{\theta}_2)}{\sqrt{\kappa^2 n_2 N^{-1}\sigma_1^2 + n_1^{-1} N\sigma_2^2}}\end{pmatrix} \to_d N\left(\begin{pmatrix}\tilde{c}_{11}\\ \tilde{c}_{12}\end{pmatrix}, \begin{pmatrix}1 & \nu_1\\ \nu_1 & 1\end{pmatrix}\right).$$

Similarly,

$$\begin{pmatrix}\frac{\hat{\theta}_2 - \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\\ \frac{\hat{\theta}_2 + \kappa\hat{\theta}_1}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\end{pmatrix} \to_d N\left(\begin{pmatrix}\tilde{c}_{21}\\ \tilde{c}_{22}\end{pmatrix}, \begin{pmatrix}1 & \nu_2\\ \nu_2 & 1\end{pmatrix}\right).$$

□

A.2 Theoretical Results for Simultaneous Test for Qualitative Interactions

In what follows, we generalize Lemma 2.2 and Proposition 2.2 to the case of unbalanced sample sizes.

Lemma A.2. *The likelihood ratio statistic $T^{\mathrm{P/N},\kappa}$ can be written as*

$$T^{\mathrm{P/N},\kappa} = \min\left\{ \frac{\left(\hat{\theta}_1 - \kappa\hat{\theta}_2\right)^2}{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}, \frac{\left(\kappa\hat{\theta}_1 - \hat{\theta}_2\right)^2}{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2} \right\} \mathbb{1}\left(\hat{\theta} \in \Theta_1^{\mathrm{P/N},\kappa}\right).$$

Proof of Lemma A.2. If $\hat{\theta}$ belongs to the null region, $\Theta_0^{\mathrm{P/N},\kappa}$, the likelihood ratio statistic is clearly zero. If $\hat{\theta}_1 > \kappa\hat{\theta}_2 > 0$ or $\hat{\theta}_1 < \kappa\hat{\theta}_2 < 0$, the likelihood ratio statistic is minimum of the distances between $\hat{\theta}$ and its projections onto the span of $(1,\kappa)^\top$ and the span of $(-\kappa,-1)^\top$. Algebra (similar to the proof of Lemma A.1) gives the desired result, and similarly if $\hat{\theta}_1 > -\kappa\hat{\theta}_2 > 0$ or $\hat{\theta}_1 < -\kappa\hat{\theta}_2 < 0$. □

Proposition A.3. *Let $n_1 + n_2 = N$, and assume $n_1/N \to \lambda > 0$. Then,*

i If θ belongs to the interior of the null region $\Theta_0^{\mathrm{P/N},\kappa}$, $T^{\mathrm{P/N},\kappa}$ converges in distribution to zero.

ii If θ is on the boundary of the null region, but $\theta \neq 0$, $\mathbb{P}(T^{\mathrm{P/N},\kappa} > t) \to \frac{1}{2}\mathbb{P}\left(\chi_1^2 > t\right)$ as $n_1, n_2 \to \infty$.

Proof of Proposition A.3. We first prove (i). If θ belongs to the interior of the null region, consistency of $\hat{\theta}$ and the continuous mapping theorem imply that $\mathbb{1}\left(\hat{\theta} \in \Theta_1^{\mathrm{P/N},\kappa}\right)$ converges in probability to zero. Then, by Slutsky's theorem, $T^{\mathrm{P/N},\kappa}$ converges in distribution to zero.

To prove *(ii)*, recall that we can write the null and alternative regions as

$$\begin{aligned}\Theta_0^{\mathrm{P/N},\kappa} =& \{0 < \kappa^{-1}\theta_1 < \theta_2 < \kappa\theta_1\} \cup \{\kappa\theta_1 < \theta_2 < \kappa^{-1}\theta_1 < 0\} \\ \Theta_1^{\mathrm{P/N},\kappa} =& \{\{\theta_1 > \kappa\theta_2\} \cap \{\theta_1 > 0\}\} \cup \{\{-\theta_1 > -\kappa\theta_2\} \cap \{\theta_1 < 0\}\} \cup \\ & \{\{\theta_2 > \kappa\theta_1\} \cap \{\theta_2 > 0\}\} \cup \{\{-\theta_2 > -\kappa\theta_1\} \cap \{\theta_2 < 0\}\}.\end{aligned}$$

Now, we write

$$\mathbb{P}\left(T^{\mathrm{P/N},\kappa} > t\right) =$$

$$\mathbb{P}\left(\frac{\left(\hat{\theta}_1 - \kappa\hat{\theta}_2\right)^2}{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > t, \frac{\left(\kappa\hat{\theta}_1 - \hat{\theta}_2\right)^2}{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > t\right) =$$

$$\begin{aligned}&\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > \sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > \sqrt{t}\right) + \\ &\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > \sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) < -\sqrt{t}\right) + \\ &\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) < -\sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) > \sqrt{t}\right) + \\ &\mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) < -\sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}}\mathbb{1}\left(\hat{\theta} \in \Theta_1\right) < -\sqrt{t}\right).\end{aligned}$$

The second and third summands above are exactly equal to 0. To see this, note that in the second term $\hat{\theta}_1 > \kappa\hat{\theta}_2$ and $\hat{\theta}_1 < \kappa^{-1}\hat{\theta}_2$ implies that $\hat{\theta} \in \Theta_0^{\mathrm{P/N},\kappa}$, and similarly for the third term. In the first and last summands, we can ignore the term $\mathbb{1}\left(\hat{\theta} \in \Theta_1^{\mathrm{P/N},\kappa}\right)$. To see this, note that in the first term $\hat{\theta}_1 > \kappa\hat{\theta}_2$ and $\hat{\theta}_1 > \kappa^{-1}\hat{\theta}_2$ imply that $\hat{\theta} \in \Theta_1^{\mathrm{P/N},\kappa}$; a similar

argument holds for the fourth term. Thus,

$$\mathbb{P}\left(T^{\mathrm{P/N},\kappa} > t\right) = \mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} > \sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} > \sqrt{t}\right) + \mathbb{P}\left(\frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} < -\sqrt{t}, \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} < -\sqrt{t}\right). \quad \text{(A.1)}$$

Suppose, for the moment, $\theta_1 = \kappa\theta_2 > 0$. Then

$$\sqrt{\frac{n_1 n_2}{N}}\begin{pmatrix}\hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2\end{pmatrix} \to_d N\left(0, \begin{pmatrix}(1-\lambda)\sigma_1^2 & 0 \\ 0 & \lambda\sigma_2^2\end{pmatrix}\right).$$

Therefore,

$$\sqrt{\frac{n_1 n_2}{N}}\begin{pmatrix}\hat{\theta}_1 - \kappa\hat{\theta}_2 \\ \kappa\hat{\theta}_1 - \hat{\theta}_2\end{pmatrix} = \sqrt{\frac{n_1 n_2}{N}}\begin{pmatrix}\hat{\theta}_1 - \kappa\hat{\theta}_2 \\ \kappa(\hat{\theta}_1 - \theta_1) - (\hat{\theta}_2 - \theta_2) + (\kappa\theta_1 - \theta_2)\end{pmatrix} \to_d N\left(0, \begin{pmatrix}(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2 & \kappa(1-\lambda)\sigma_1^2 + \kappa\lambda\sigma_2^2 \\ \kappa(1-\lambda)\sigma_1^2 + \kappa\lambda\sigma_2^2 & \kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2\end{pmatrix}\right) + \begin{pmatrix}0 \\ \infty\end{pmatrix}.$$

Thus,

$$\mathbb{P}\left(T^{\mathrm{P/N},\kappa} > t\right) \to 1 - \Phi\left(\sqrt{t}\right) = \frac{1}{2}\mathbb{P}\left(\chi_1^2 > t\right).$$

A similar argument follows if $\theta_1 = \kappa\theta_2 < 0$, $\theta_2 = \kappa\theta_1 > 0$, or $\theta_2 = \kappa\theta_1 < 0$. □

In Proposition A.4, we characterize the local asymptotic behavior of the omnibus test for qualitative interactions.

Proposition A.4. *Let $n_1 + n_2 = N$, and assume $n_1/N \to \lambda > 0$ as $n_1, n_2 \to \infty$. Suppose $\theta_1 = n_1^{-1/2}c_1, \theta_2 n^{-1/2}c_2$ with $c_1, c_2 \geq 0$. Further, assume $\hat{\theta}$ is locally regular in the sense that $\sqrt{n_g}\hat{\theta}_g \to_d N(c_g, \sigma_g^2)$ for $g \in \{1, 2\}$. Then as $n_1, n_2 \to \infty$, for all $t > 0$, $\mathbb{P}\left(T^{\mathrm{P/N},\kappa} > t\right)$ converges to*

$$\mathbb{P}\left(V_1 > \sqrt{t}, V_2 > \sqrt{t}\right) + \mathbb{P}\left(V_1 < -\sqrt{t}, V_2 < -\sqrt{t}\right),$$

where (V_1, V_2) *follows a bivariate normal distribution with mean* $(\tilde{c}_1, \tilde{c}_2)$*, unit variance, and correlation* ν*, where*

$$\tilde{c}_1 = \frac{(1-\lambda)c_1 - \kappa\lambda c_2}{\sqrt{(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2}}$$
$$\tilde{c}_2 = \frac{\kappa(1-\lambda)c_1 - \lambda c_2}{\sqrt{\kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}}$$
$$\nu = \frac{\kappa(1-\lambda)\sigma_1^2 + \kappa\lambda\sigma_2^2}{\sqrt{(1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2}\sqrt{\kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}}.$$

Remark A.2. *Setting* $c_1 = c_2 = 0$*, we obtain the limiting distribution of* $T^{P/N,\kappa}$ *when* $\theta_1 = \theta_2 = 0$*, which is critical for calculating the p-value* $\rho^{P/N,\kappa}(t)$*.*

Proof of Proposition A.4. First,

$$\sqrt{\frac{n_1 n_2}{N}} \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \to_d N\left(\begin{pmatrix} (1-\lambda)c_1 \\ \lambda c_2 \end{pmatrix}, \begin{pmatrix} (1-\lambda)\sigma_1^2 & 0 \\ 0 & \lambda\sigma_2^2 \end{pmatrix}. \right)$$

Also,

$$\sqrt{\frac{n_1 n_2}{N}} \begin{pmatrix} 1 & -\kappa \\ \kappa & -1 \end{pmatrix} \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \to_d N\left(\begin{pmatrix} (1-\lambda)c_1 - \kappa\lambda c_2 \\ \kappa(1-\lambda)c_1 - \lambda c_2 \end{pmatrix}, \begin{pmatrix} (1-\lambda)\sigma_1^2 + \kappa^2\lambda\sigma_2^2 & \kappa(1-\lambda)\sigma_1^2 + \kappa\lambda\sigma_2^2 \\ \kappa(1-\lambda)\sigma_1^2 + \kappa\lambda\sigma_2^2 & \kappa^2(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2 \end{pmatrix} \right).$$

Thus,

$$\begin{pmatrix} \frac{\hat{\theta}_1 - \kappa\hat{\theta}_2}{\sqrt{n_1^{-1}\sigma_1^2 + \kappa^2 n_2^{-1}\sigma_2^2}} \\ \frac{\kappa\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\kappa^2 n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{\frac{n_1 n_2}{N}}(\hat{\theta}_1 - \kappa\hat{\theta}_2)}{\sqrt{n_2 N^{-1}\sigma_1^2 + \kappa^2 n_1 N^{-1}\sigma_2^2}} \\ \frac{\sqrt{\frac{n_1 n_2}{N}}(\kappa\hat{\theta}_1 - \hat{\theta}_2)}{\sqrt{\kappa^2 n_2 N^{-1}\sigma_1^2 + n_1^{-1} N\sigma_2^2}} \end{pmatrix} \to_d N\left(\begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix}, \begin{pmatrix} 1 & \nu \\ \nu & 1 \end{pmatrix} \right).$$

Application of (A.1) completes the argument. □

A.3 Calculation of $\kappa_{\max}^{\alpha}$

Define t^{κ} as a realization of the likelihood ratio statistic T^{κ} for the relative difference hypothesis, calculated on the observed data. For a pre-specified $\alpha < 1$, recall that $\kappa_{\max}^{\alpha} \equiv \sup\{\kappa : \kappa > 1, \rho^{\kappa}(t^{\kappa}) < \alpha\}$ is defined as the largest $\kappa > 1$ such that the likelihood

ratio test rejects the relative difference null. Defining $F^{\kappa}(t) \equiv \lim_{n_1,n_2\to\infty} \mathbb{P}\left(T^{\kappa} > t|\theta = 0\right)$, we can evaluate $\kappa^{\alpha}_{\max}$ as

$$\begin{aligned}\kappa^{\alpha}_{\max} &= \min\left\{\sup\left\{\kappa : \kappa > 1, 1 - \Phi\left(t^{\kappa}\right) < \alpha\right\}, \sup\{\kappa : \kappa > 1, 1 - F^{\kappa}(t^{\kappa}) < \alpha\}\right\} \\ &\equiv \min\{\pi_1, \pi_2\}.\end{aligned}$$

Therefore, we are only required to calculate π_1 and π_2.

Both π_1 and π_2 can be calculated with root-finding algorithms. To see this, we first observe that t^{κ} is monotone in κ, recalling that the numerator of $T^{\kappa} = \frac{\hat{\theta}_{\max} - \kappa\hat{\theta}_{\min}}{\hat{\tau}_{\max} + \kappa^2\hat{\tau}_{\min}}$ decreases in κ while the denominator increases. Monotonicity of $\Phi(\cdot)$ implies that $1-\Phi(t^{\kappa})$ is monotone in κ. Now, applying Proposition A.2,

$$1 - F^{\kappa}(t) = 2\left\{\mathbb{P}\left(W_{11} > t, W_{12} > t\right) + \mathbb{P}\left(W_{21} > t, W_{22} > t\right)\right\},$$

where (W_{11}, W_{12}) and (W_{21}, W_{22}) follow bivariate normal distributions with mean zero, unit variance, and correlations ν_1 and ν_2, as defined in Proposition A.2, respectively. Both ν_1 and ν_2 are monotone decreasing in κ, so Theorem 8 in Müller (2001) implies that $1 - F^{\kappa}(t)$ is monotone in t. Thus $1 - F^{\kappa}(t^{\kappa})$ is monotone in κ.

Monotonicity of $1 - \Phi(t^{\kappa})$ and $1 - F^{\kappa}(t^{\kappa})$ implies that if the likelihood ratio test rejects the null hypothesis for some $\kappa > 1$, π_1 and π_2 are the unique roots of

$$\begin{aligned}f_1(\kappa) &= 1 - \Phi(t^{\kappa}) - \alpha \\ f_2(\kappa) &= 1 - F^{\kappa}(t^{\kappa}) - \alpha,\end{aligned}$$

respectively. These roots can be easily calculated via, e.g., the bisection method, which is implemented in the `uniroot` function in R.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 De-biased Group LASSO Estimator

In this subsection, we derive a de-biased group LASSO estimator. Our construction is essentially the same as the one presented in van de Geer (2016).

With $\mathcal{V}_j$ as defined in (3.10), let $\mathcal{V}^g_{-j} = \left(\mathcal{V}^g_1, \cdots, \mathcal{V}^g_{j-1}, \mathcal{V}^g_{j+1}, \cdots, \mathcal{V}^g_p\right)$ be an $n \times (p-1)d$ dimensional matrix. For $\alpha_{j,1}, \ldots, \alpha_{j,p} \in \mathbb{R}^d$, let $\boldsymbol{\alpha}_j = \left(\alpha_{j,1}^\top, \cdots, \alpha_{j,p}^\top\right)^\top$, let $\mathcal{P}_j\left(\boldsymbol{\alpha}_j\right) = \sum_{k \neq j} \|\alpha_{j,k}\|_2$, and let $\nabla \mathcal{P}_j$ denote the sub-gradient of $\mathcal{P}_j$. We can express the sub-gradient as $\nabla \mathcal{P}_j(\boldsymbol{\alpha}_j) = \left((\nabla\|\alpha_{j,1}\|_2)^\top, \cdots, (\nabla\|\alpha_{j,p}\|_2)^\top\right)^\top$ where $\nabla\|\alpha_{j,k}\|_2 = \alpha_{j,k}/\|\alpha_{j,k}\|_2$ if $\|\alpha_{j,k}\|_2 \neq 0$, and $\nabla\|\alpha_{j,k}\|_2$ is otherwise a vector with ℓ_2 norm less than one. The KKT conditions for the group LASSO imply that the estimate $\tilde{\boldsymbol{\alpha}}^g_j$ satisfies

$$(n^g)^{-1}\left(\mathcal{V}^g_{-j}\right)^\top\left(\mathbf{X}^g_j - \mathcal{V}^g_{-j}\tilde{\boldsymbol{\alpha}}^g_j\right) = -\lambda \nabla \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}^g_j\right).$$

With some algebra, we can rewrite this as

$$(n^g)^{-1}\left(\mathcal{V}^g_{-j}\right)^\top \mathcal{V}^g_{-j}\left(\tilde{\boldsymbol{\alpha}}^g_j - \boldsymbol{\alpha}^{g,*}_j\right) = -\lambda \nabla \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}^g_j\right) + \left(\mathcal{V}^g_{-j}\right)^\top\left(\mathbf{X}^g_j - \mathcal{V}^g_{-j}\boldsymbol{\alpha}^{g,*}_j\right).$$

Let Σ_j be defined as the matrix

$$\Sigma_j = \mathbb{E}\left[(n^g)^{-1}\left(\mathcal{V}^g_{-j}\right)^\top \mathcal{V}^g_{-j}\right],$$

and let $\tilde{M}_j$ be an estimate of Σ_j^{-1}. We can write $\left(\tilde{\boldsymbol{\alpha}}^g_j - \tilde{\boldsymbol{\alpha}}^{g,*}_j\right)$ as

$$\begin{aligned}\left(\tilde{\boldsymbol{\alpha}}^g_j - \boldsymbol{\alpha}^{g,*}_j\right) =& \underbrace{-\lambda \tilde{M}_j \nabla \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}^g_j\right)}_{\text{(i)}} + \underbrace{(n^g)^{-1}\tilde{M}_j\left(\mathcal{V}^g_{-j}\right)^\top\left(\mathbf{X}^g_j - \mathcal{V}^g_{-j}\boldsymbol{\alpha}^{g,*}_j\right)}_{\text{(ii)}} + \\ & \underbrace{\left\{I - (n^g)^{-1}\tilde{M}_j\left(\mathcal{V}^g_{-j}\right)^\top \mathcal{V}^g_{-j}\right\}\left(\tilde{\boldsymbol{\alpha}}^g_j - \boldsymbol{\alpha}^{g,*}_j\right)}_{\text{(iii)}}. \end{aligned} \tag{B.1}$$

The first term (i) in (B.1) is an approximation for the bias of the group LASSO estimate. This term is a function only of the observed data and not of any unknown quantities. This term can therefore be directly added to the initial estimate $\tilde{\boldsymbol{\alpha}}_j^g$. If $\tilde{M}_j$ is a consistent estimate of Σ_j^{-1}, the second term (ii) is asymptotically equivalent to

$$\Sigma_j^{-1} \left(\mathcal{V}_{-j}^g\right)^\top \left(\mathbf{X}_j^g - \mathcal{V}_{-j}^g \boldsymbol{\alpha}_j^{g,*}\right).$$

Thus, (ii) is asymptotically equivalent to a sample average of mean zero *i.i.d.* random variables. The central limit theorem can then be applied to establish convergence in distribution to the multivariate normal distribution at an $n^{1/2}$ rate for any low-dimensional sub-vector. The third term will also be asymptotically negligible if $\tilde{M}_j$ is an approximate inverse of $(n^g)^{-1} \left(\mathcal{V}_{-j}^g\right)^\top \mathcal{V}_{-j}^g$. This would suggest that an estimator of the form

$$\check{\boldsymbol{\alpha}}_j^g = \tilde{\boldsymbol{\alpha}}_j^g + \lambda \tilde{M}_j \nabla \mathcal{P}_j \left(\tilde{\boldsymbol{\alpha}}_j^g\right)$$

will be asymptotically normal for an appropriate choice of $\tilde{M}_j$.

Before describing our construction of $\tilde{M}_j$, we find it helpful to consider an alternative expression for Σ_j^{-1}. We define the $d \times d$ matrices $\Gamma_{j,k,l}^*$ as

$$\Gamma_{j,k,1}^*, \ldots, \Gamma_{j,k,p}^* = \underset{\Gamma_1,\ldots,\Gamma_p \in \mathbb{R}^{d\times d}}{\arg\min}\ \mathbb{E}\left[\text{trace}\left\{(n^g)^{-1}\left(\mathcal{V}_k^g - \sum_{l\neq k,j} \mathcal{V}_l^g \Gamma_l\right)^\top \left(\mathcal{V}_k^g - \sum_{l\neq k,j} \mathcal{V}_l^g \Gamma_l\right)\right\}\right].$$

We also define the $d \times d$ matrix $\tilde{C}_{j,k}$ as

$$C_{j,k}^* = \mathbb{E}\left[(n^g)^{-1}\left(\mathcal{V}_k^g - \sum_{l\neq k,j} \mathcal{V}_l^g \Gamma_{j,k,l}^*\right)^\top \mathcal{V}_k^g\right].$$

It can be shown that Σ_j^{-1} can be expressed as

$$\Sigma_j^{-1} = \begin{pmatrix} \left(C_{j,1}^*\right)^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \left(C_{j,p}^*\right)^{-1} \end{pmatrix} \begin{pmatrix} I & -\Gamma_{j,1,2}^* & \cdots & -\Gamma_{j,1,p}^* \\ -\Gamma_{j,2,1}^* & I & \cdots & -\Gamma_{j,2,p}^* \\ \vdots & \vdots & \ddots & \vdots \\ -\Gamma_{j,p,1}^* & -\Gamma_{j,p,2}^* & \cdots & I \end{pmatrix}.$$

We can thus estimate Σ_j^{-1} by performing a series of regressions to estimate each matrix $\Gamma^*_{j,k,l}$.

Following the approach of van de Geer et al. (2014), we use a group LASSO variant of the nodewise LASSO to construct $\tilde{M}_j$. To proceed, we require some additional notation. For any $d \times d$ matrix $\Gamma = (\gamma_1, \cdots, \gamma_d)$ for $d-$dimensional vectors γ_c, let $\|\Gamma\|_{2,*} = \sum_{c=1}^d \|\gamma_c\|_2$. Let $\nabla\|\Gamma\|_{2,*} = (\gamma_1/\|\gamma_1\|_2, \cdots, \gamma_d/\|\gamma_d\|_2)$ be the subgradient of $\|\Gamma\|_{2,*}$. We use the group LASSO to obtain estimates $\tilde{\Gamma}_{j,k,l}$ of $\Gamma^*_{j,k,l}$:

$$\tilde{\Gamma}_{j,k,1}, \ldots, \tilde{\Gamma}_{j,k,p} = \\ \underset{\Gamma_1,\ldots,\Gamma_p \in \mathbb{R}^{d\times d}}{\arg\min} \operatorname{trace}\left\{ (n^g)^{-1} \left(\mathcal{V}_k^g - \sum_{l \neq k,j} \mathcal{V}_l^g \Gamma_l \right)^\top \left(\mathcal{V}_k^g - \sum_{l \neq k,j} \mathcal{V}_l^g \Gamma_l \right) \right\} + \omega \sum_{l \neq k,j} \|\Gamma_l\|_{2,*}. \quad \text{(B.2)}$$

We then estimate $C^*_{j,k}$ as

$$\tilde{C}_{j,k} = (n^g)^{-1} \left(\mathcal{V}_k^g - \sum_{l \neq k,j} \mathcal{V}_l^g \tilde{\Gamma}_{j,k,l} \right)^\top (\mathcal{V}_k^g)\,.$$

Our estimate $\tilde{M}_j$ takes the form

$$\tilde{M}_j = \begin{pmatrix} \tilde{C}_{j,1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \tilde{C}_{j,p}^{-1} \end{pmatrix} \begin{pmatrix} I & -\tilde{\Gamma}_{j,1,2} & \cdots & -\tilde{\Gamma}_{j,1,p} \\ -\tilde{\Gamma}_{j,2,1} & I & \cdots & -\tilde{\Gamma}_{j,2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\Gamma}_{j,p,1} & -\tilde{\Gamma}_{j,p,2} & \cdots & I \end{pmatrix}.$$

With this construction of $\tilde{M}_j$, we can establish a bound on the remainder term (iii) in (B.1). To show this, we make use of the following lemma, which states a special case of the dual norm inequality for the group LASSO norm $\mathcal{P}_j$ (see, e.g., Chapter 6 of van de Geer, 2016).

Lemma B.1. *Let $a_1, \ldots, a_p$ and $b_1, \ldots, b_p$ be d-dimensional vectors, and let $\mathbf{a} = \left(a_1^\top, \cdots, a_p^\top\right)^\top$ and $\mathbf{b} = \left(b_1^\top, \cdots, b_p^\top\right)^\top$ be pd-dimensional vectors. Then*

$$\langle \mathbf{a}, \mathbf{b} \rangle \leq \left(\sum_{j=1}^p \|a_j\|_2 \right) \max_j \|b_j\|_2\,.$$

The KKT conditions for (B.2) imply that for all $l \neq j, k$

$$(n^g)^{-1} \left(\mathcal{V}_l^g\right)^\top \left(\mathcal{V}_k^g - \sum_{r \neq k,j} \mathcal{V}_r^g \tilde{\Gamma}_{j,k,r} \right) = -\omega \nabla \left\| \tilde{\Gamma}_{j,k,l} \right\|_{2,*}. \tag{B.3}$$

Lemma B.1 and (B.3) imply that

$$\left\| \begin{pmatrix} \tilde{C}_{j,1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \tilde{C}_{j,p} \end{pmatrix} \left\{ I - (n^g)^{-1} \tilde{M}_j \left(\mathcal{V}_{-j}^g\right)^\top \mathcal{V}_{-j}^g \right\} \left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \right) \right\|_\infty \leq \omega \mathcal{P}_j \left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \right),$$

where $\| \cdot \|_\infty$ is the ℓ_∞ norm. With $\omega \propto \{\log(p)/n\}^{1/2}$, $\tilde{M}_j$ can be shown to be consistent under sparsity of $\Gamma_{j,k,l}^*$ (i.e., only a few matrices $\Gamma_{j,k,l}^*$ have some nonzero columns) and some additional regularity conditions. Additionally, it can be shown under sparsity of $\boldsymbol{\alpha}^{g,*}$ (i.e., very few vectors $\alpha_{j,k}^{g,*}$ are nonzero) and some additional regularity conditions that $\mathcal{P}_j \left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \right) = O_P \left(\{\log(p)/n\}^{1/2} \right)$. Thus, a scaled version of the remainder term (iii) is $o_P(n^{-1/2})$ if $n^{-1/2} \log(p) \to 0$. We refer readers to Chapter 8 of Bühlmann and van de Geer (2011) for a more comprehensive discussion of assumptions required for consistency of the group LASSO.

We now express the de-biased group LASSO estimator for $\alpha_{j,k}^{g,*}$ as

$$\check{\alpha}_{j,k}^g = \tilde{\alpha}_{j,k}^g + (n^g)^{-1} \tilde{C}_{j,k}^{-1} \left(\mathcal{V}_k^g - \sum_{l \neq j,k} \tilde{\Gamma}_{j,k,l} \mathcal{V}_l^g \right)^\top \left(\mathbf{X}_j^g - \mathcal{V}_{-j}^g \tilde{\boldsymbol{\alpha}}_j^g \right).$$

We have established that $\check{\alpha}_{j,k}^g$ can be written as

$$\tilde{C}_{j,k} \left(\check{\alpha}_{j,k}^g - \alpha_{j,k}^{g,*} \right) = (n^g)^{-1} \left(\mathcal{V}_k^g - \sum_{l \neq j,k} \Gamma_{j,k,l}^* \mathcal{V}_l^g \right)^\top \left(\mathbf{X}_j^g - \mathcal{V}_{-j}^g \boldsymbol{\alpha}_j^{g,*} \right) + o_P(n^{-1/2}).$$

As stated above, the central limit theorem implies asymptotic normality of $\check{\alpha}_{j,k}^g$.

We now construct an estimate for the variance of $\check{\alpha}_{j,k}^g$. Suppose the residual $\mathbf{X}_j^g - \mathcal{V}_{-j}^g \boldsymbol{\alpha}_j^{g,*}$ is independent of $\mathcal{V}^g$, and let τ_j^g denote the residual variance

$$\tau_j^g = \mathbb{E} \left[\left(X_j^g - \sum_{k \neq j} \left\langle \phi_i^g, \alpha_{j,k}^{g,*} \right\rangle X_k^g \right)^2 \right].$$

We can approximate the variance of $\check{\alpha}^g_{j,k}$ as

$$\check{\Omega}^g_{j,k} = (n^g)^{-2}\,\tau^g_j \tilde{C}^{-1}_{j,k}\left(\mathcal{V}^g_k - \sum_{l\neq j,k}\tilde{\Gamma}_{j,k,l}\mathcal{V}^g_l\right)^\top\left(\mathcal{V}^g_k - \sum_{l\neq j,k}\tilde{\Gamma}_{j,k,l}\mathcal{V}^g_l\right)\left(\tilde{C}^{-1}_{j,k}\right)^\top.$$

As τ^g_j is typically unknown, we instead us the estimate

$$\tilde{\tau}^g_j = \frac{\left\|\mathbf{X}^g_j - \mathcal{V}^g_{-j}\tilde{\boldsymbol{\alpha}}^g_j\right\|_2^2}{n - \widehat{df}},$$

where $\widehat{df}$ is an estimate of the degrees of freedom for the group LASSO estimate $\tilde{\boldsymbol{\alpha}}^g_j$. In our implementation, we use the estimate proposed by Breheny and Huang (2009). Let $\tilde{\alpha}^g_{j,k,l}$ be the l-th element of $\tilde{\alpha}^g_{j,k}$, and let $\mathcal{V}^g_{k,l}$ denote the l-th column of $\mathcal{V}^g_k$. We then define

$$\bar{\alpha}^g_{j,k,l} = \frac{\langle \mathbf{X}^g_j - \mathcal{V}^g_{-j}\tilde{\boldsymbol{\alpha}}^g_j + \mathcal{V}^g_{k,l}\tilde{\alpha}^g_{j,k,l}, \mathcal{V}^g_{k,l}\rangle}{\langle \mathcal{V}^g_{k,l}, \mathcal{V}^g_{k,l}\rangle},$$

and estimate the degrees of freedom as

$$\hat{df} = \sum_{k\neq j}\sum_{l=1}^{d}\frac{\tilde{\alpha}^g_{j,k,l}}{\bar{\alpha}^g_{j,k,l}}.$$

B.2 Generalized Score Matching Estimator

In this section, we establish consistency of the regularized score matching estimator and derive a bias-corrected estimator.

B.2.1 Form of Generalized Score Matching Loss

Below, we restate Theorem 3 of Yu et al. (2019), which provides conditions under which the score matching loss (3.15) can be expressed as (3.16).

Theorem B.1. *Assume the following conditions hold:*

$$\lim_{z_j \to \infty} h^*(z)(z_j) \left\{ \frac{\partial}{\partial z_j} h(z_j) \right\} = 0 \quad \forall\, z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_p \in \mathbb{R}_+, \quad \forall h \in \mathcal{H}$$

$$\lim_{z_j \to 0} h^*(z)(z_j) \left\{ \frac{\partial}{\partial z_j} h(z_j) \right\} = 0 \quad \forall\, z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_p \in \mathbb{R}_+, \quad \forall h \in \mathcal{H}$$

$$\sup_{h \in \mathcal{H}} \int \left\| \nabla \log h(z) \circ v^{1/2}(z) \right\|_2^2 h^*(z) dz < \infty$$

$$\sup_{h \in \mathcal{H}} \int \left\| \left\{ \nabla \log h(z) \circ v^{1/2}(z) \right\}' \right\|_1 h^*(z) dz < \infty,$$

where the prime symbol denotes the element-wise derivative. Then (3.15) *and* (3.16) *are equivalent up to an additive constant that does not depend on* h*.*

B.2.2 Generalized Score Matching Estimator in Low Dimensions

In this section, we provide an explicit form for the generalized score matching estimator in the low-dimensional setting and state its limiting distribution. We first introduce some additional notation below that allows for the generalized score matching loss to be written in a condensed form. Recall the form of the conditional density for the pairwise interaction model in (3.17). We define

$$\mathcal{V}^g_{j,k,1} = \begin{pmatrix} v_j^{1/2}\left(X^g_{1,j}\right) \dot{\psi}\left(X^g_{1,j}, X^g_{1,k}\right) \times \phi\left(W^g_1\right) \\ \vdots \\ v_j^{1/2}\left(X^g_{n^g,j}\right) \dot{\psi}\left(X^g_{n^g,j}, X^g_{n^g,k}\right) \times \phi\left(W^g_{n^g}\right) \end{pmatrix},$$

$$\mathcal{V}^g_{2,j} = \begin{pmatrix} v_j^{1/2}\left(X^g_{1,j}\right) \times \left\{ \dot{\zeta}\left(X^g_{1,j}, \phi_1(W^g_1)\right), \cdots, \dot{\zeta}\left(X^g_{1,j}, \phi_d(W^g_1)\right) \right\} \\ \vdots \\ v_j^{1/2}\left(X^g_{n^g,j}\right) \times \left\{ \dot{\zeta}\left(X^g_{n^g,j}, \phi_1(W^g_{n^g})\right), \cdots, \dot{\zeta}\left(X^g_{n^g,j}, \phi_d(W^g_{n^g})\right) \right\} \end{pmatrix},$$

$$\mathcal{U}^g_{j,k,1} = \begin{pmatrix} \left\{\dot{v}_j\left(X^g_{1,j}\right)\dot{\psi}\left(X^g_{1,j}, X^g_{1,k}\right) + v_j\left(X^g_{1,j}\right)\ddot{\psi}\left(X^g_{1,j}, X^g_{1,k}\right)\right\} \times \phi\left(W^g_1\right) \\ \vdots \\ \left\{\dot{v}_j\left(X^g_{1,j}\right)\dot{\psi}\left(X^g_{n^g,j}, X^g_{n^g,k}\right) + v_j\left(X^g_{n^g,j}\right)\ddot{\psi}\left(X^g_{1,j}, X^g_{n^g,k}\right)\right\} \times \phi\left(W^g_{n^g}\right) \end{pmatrix},$$

$$\mathcal{U}^g_{j,2} = \begin{pmatrix} v_j\left(X^g_{1,j}\right)\ddot{\zeta}\left(X^g_{1,j}, \phi_1(W^g_1)\right) & \cdots & v_j\left(X^g_{1,j}\right)\ddot{\zeta}\left(X^g_{1,j}, \phi_d(W^g_1)\right) \\ \vdots & \ddots & \vdots \\ v_j\left(X^g_{n^g,j}\right)\ddot{\zeta}\left(X^g_{n^g,j}, \phi_1(W^g_{n^g})\right) & \cdots & v_j\left(X^g_{n^g,j}\right)\ddot{\zeta}\left(X^g_{n^g,j}, \phi_d(W^g_{n^g})\right) \end{pmatrix} + \begin{pmatrix} \dot{v}_j\left(X^g_{1,j}\right)\dot{\zeta}\left(X^g_{1,j}, \phi_1(W^g_1)\right) & \cdots & \dot{v}_j\left(X^g_{1,j}\right)\dot{\zeta}\left(X^g_{1,j}, \phi_d(W^g_1)\right) \\ \vdots & \ddots & \vdots \\ \dot{v}_j\left(X^g_{n^g,j}\right)\dot{\zeta}\left(X^g_{n^g,j}, \phi_1(W^g_{n^g})\right) & \cdots & \dot{v}_j\left(X^g_{n^g,j}\right)\dot{\zeta}\left(X^g_{n^g,j}, \phi_d(W^g_{n^g})\right) \end{pmatrix},$$

$$\mathcal{V}^g_{j,1} = \begin{pmatrix} \mathcal{V}^g_{j,1,1} \\ \vdots \\ \mathcal{V}^g_{j,p,1} \end{pmatrix}; \quad \mathcal{U}^g_{j,1} = \begin{pmatrix} \mathcal{U}^g_{1,j,1} \\ \vdots \\ \mathcal{U}^g_{j,p,1} \end{pmatrix}.$$

Let $\boldsymbol{\alpha}_j = \left(\alpha^\top_{j,1}, \cdots, \alpha^\top_{j,p}\right)^\top$ for $\alpha_{j,k} \in \mathbb{R}^d$ and $\boldsymbol{\theta}_j = (\theta_{j,1}, \cdots, \theta_{j,d})^\top$ for $\theta_{j,c} \in \mathbb{R}$. We can express the empirical score matching loss (3.18) as

$$L^g_{n,j}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j) = (2n^g)^{-1}\left(\mathcal{V}^g_{j,1}\boldsymbol{\alpha}_j + \mathcal{V}^g_{2,j}\boldsymbol{\theta}_j\right)^\top\left(\mathcal{V}^g_{j,1}\boldsymbol{\alpha}_j + \mathcal{V}^g_{2,j}\boldsymbol{\theta}_j\right) + (n^g)^{-1}\mathbf{1}^\top\left(\mathcal{U}^g_{1,j}\boldsymbol{\alpha}_j + \mathcal{U}^g_{2,j}\boldsymbol{\theta}_j\right).$$

We write the gradient of the risk function as

$$\nabla L^g_{n,j}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j) = (n^g)^{-1}\begin{pmatrix} \left(\mathcal{V}^g_{j,1}\right)^\top\mathcal{V}^g_{j,1} & \left(\mathcal{V}^g_{j,1}\right)^\top\mathcal{V}^g_{j,2} \\ \left(\mathcal{V}^g_{j,2}\right)^\top\mathcal{V}^g_{j,1} & \left(\mathcal{V}^g_{j,2}\right)^\top\mathcal{V}^g_{j,2} \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha}_j \\ \boldsymbol{\theta}_j \end{pmatrix} + (n^g)^{-1}\begin{pmatrix} \left(\mathcal{U}^g_{j,1}\right)^\top\mathbf{1} \\ \left(\mathcal{U}^g_{j,2}\right)^\top\mathbf{1} \end{pmatrix}.$$

Thus, the minimizer $(\hat{\boldsymbol{\alpha}}^g_j, \hat{\boldsymbol{\theta}}^g_j)$ of the empirical loss takes the form

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}^g_j \\ \hat{\boldsymbol{\theta}}^g_j \end{pmatrix} = -\begin{pmatrix} \left(\mathcal{V}^g_{j,1}\right)^\top\mathcal{V}^g_{j,1} & \left(\mathcal{V}^g_{j,1}\right)^\top\mathcal{V}^g_{j,2} \\ \left(\mathcal{V}^g_{j,2}\right)^\top\mathcal{V}^g_{j,1} & \left(\mathcal{V}^g_{j,2}\right)^\top\mathcal{V}^g_{j,2} \end{pmatrix}^{-1}\begin{pmatrix} \left(\mathcal{U}^g_{j,1}\right)^\top\mathbf{1} \\ \left(\mathcal{U}^g_{j,2}\right)^\top\mathbf{1} \end{pmatrix}.$$

By applying Theorem 5.23 of van der Vaart (2000),

$$(n^g)^{1/2} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \hat{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*} \end{pmatrix} \to_d N\left(0, \Big(ABA\Big)\right),$$

where the matrices A and B are defined as

$$A = E\left[(n^g)^{-1} \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g \end{pmatrix}\right]^{-1},$$

$$B = \mathrm{Cov}\left((n^g)^{-1} \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_j^{g,*} \\ \boldsymbol{\theta}_j^{g,*} \end{pmatrix} + (n^g)^{-1} \begin{pmatrix} \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \\ \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1} \end{pmatrix}\right).$$

We estimate the variance of $(\hat{\boldsymbol{\alpha}}_j^g, \hat{\boldsymbol{\theta}}_j^g)$ as $\hat{\Omega}_j^g = (n^g)^{-1} \hat{A}\hat{B}\hat{A}$, where

$$\hat{A} = n^g \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g \end{pmatrix}^{-1},$$

$$\hat{B} = (n^g)^{-1} \hat{\xi}^\top \hat{\xi}, \quad \hat{\xi} = \begin{pmatrix} \mathrm{diag}\left(\mathcal{V}_{j,1}^g \hat{\boldsymbol{\alpha}}_j^g + \mathcal{V}_{j,2}^g \hat{\boldsymbol{\theta}}_j^g\right) \mathcal{V}_{j,1}^g \\ \mathrm{diag}\left(\mathcal{V}_{j,1}^g \hat{\boldsymbol{\alpha}}_j^g + \mathcal{V}_{j,2}^g \hat{\boldsymbol{\theta}}_j^g\right) \mathcal{V}_{j,2}^g \end{pmatrix} + \begin{pmatrix} \mathcal{U}_{j,1}^g \\ \mathcal{U}_{j,2}^g \end{pmatrix}.$$

B.2.3 Consistency of Regularized Generalized Score Matching Estimator

In this subsection, we argue that the regularized generalized score matching estimators $\tilde{\boldsymbol{\alpha}}_j^g$ and $\tilde{\boldsymbol{\theta}}_j^g$ from (3.19) are consistent. Let $\mathcal{P}_j(\boldsymbol{\alpha}_j) = \sum_{j=1}^p \|\alpha_{j,k}\|_2$. We establish convergence rates of $\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)$ and $\left\|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right\|_2$. Our approach is based on proof techniques described in Chapter 6 of Bühlmann and van de Geer (2011).

Our result requires a notion of compatibility between the penalty function $\mathcal{P}_j$ and the loss $L_{n,j}^g$. Such notions are commonly assumed in the high-dimensional literature. Below, we define the compatibility condition.

Definition B.1 (Compatibility Condition)**.** *Let S be a set containing indices of the nonzero elements of $\boldsymbol{\alpha}_j^{g,*}$, and let $\bar{S}$ denote the complement of S. Let $\mathbb{1}_S$ be a $(p-1)d$-dimensional vector where the r-th element is one if $r \in S$, and zero otherwise. The group LASSO*

compatibility condition holds for the index set $S \subset \{1, \ldots, p\}$ *and for constant* $C > 0$ *if for all* $\Omega\left(\boldsymbol{\alpha}_j \circ \mathbb{1}_S\right) \leq 3\Omega\left(\boldsymbol{\alpha}_j \circ \mathbb{1}_{\bar{S}}\right) + \|\boldsymbol{\theta}_j\|_2$,

$$\Omega\left(\boldsymbol{\alpha}_j \circ \mathbb{1}_S\right) + \|\boldsymbol{\theta}_j\|/2 \leq \frac{|S|^{1/2}}{C}\left\{\begin{pmatrix}\boldsymbol{\alpha}_j^\top & \boldsymbol{\theta}_j^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_j \\ \boldsymbol{\theta}_j\end{pmatrix}\right\}^{1/2},$$

where $\circ$ *is the element-wise product operator.*

Theorem B.2. *Let* $\mathcal{E}$ *be the set*

$$\begin{aligned}\mathcal{E} = &\left\{\max_{k \neq j}\left\{\left\|\left(\mathcal{V}_{j,k,1}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1}\right\|_2\right\} \leq n^g\lambda_0\right\} \cap \\ &\left\{\left\|\left(\mathcal{V}_{j,k,2}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1}\right\|_2 \leq n^g\lambda_0\right\}\end{aligned}$$

for some $\lambda_0 \leq \lambda/2$. *Suppose the compatibility condition also holds. Then on the set* $\mathcal{E}$,

$$\mathcal{P}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) + \|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\|_2 \leq \frac{\lambda 4|S|}{C^2}.$$

Proof of Theorem B.2. The regularized score matching estimator $\tilde{\boldsymbol{\alpha}}_j^g$ necessarily satisfies the following basic inequality:

$$L_{n,j}^g\left(\tilde{\boldsymbol{\alpha}}_j^g, \tilde{\boldsymbol{\theta}}_j^g\right) + \lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g\right) \leq L_{n,j}^g\left(\boldsymbol{\alpha}_j^{g,*}, \boldsymbol{\theta}_j^{g,*}\right) + \lambda\mathcal{P}_j\left(\boldsymbol{\alpha}_j^{g,*}\right).$$

With some algebra, this inequality can be rewritten as

$$\begin{aligned}&(2n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix} + \lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g\right) \leq \\ &-(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \\ \left(\mathcal{V}_{j,2}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1}\end{pmatrix} + \lambda\mathcal{P}_j\left(\boldsymbol{\alpha}_j^{g,*}\right).\end{aligned}$$

By Lemma B.1, on the set $\mathcal{E}$ and using $\lambda \geq \lambda_0/2$ we get

$$\begin{aligned}&(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix} + 2\lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g\right) \leq \\ &\lambda\left\|\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\right\|_2 + 2\lambda\mathcal{P}_j\left(\boldsymbol{\alpha}_j^{g,*}\right) + \lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right).\end{aligned}$$

On the left hand side, we apply the triangle inequality to get

$$\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g\right) = \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_S\right) + \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_{\bar{S}}\right) \geq \mathcal{P}_j\left(\boldsymbol{\alpha}_j^{g,*} \circ \mathbb{1}_S\right) - \mathcal{P}_j\left(\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) \circ \mathbb{1}_S\right) + \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_{\bar{S}}\right).$$

On the right hand side, we observe that

$$\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) = \mathcal{P}_j\left(\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) \circ \mathbb{1}_S\right) + \mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_{\bar{S}}\right).$$

We then have

$$(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix} + \lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_{\bar{S}}\right) \leq$$
$$\lambda\left\|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right\|_2 + 3\lambda\mathcal{P}_j\left(\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) \circ \mathbb{1}_S\right).$$

Now,

$$(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix} +$$
$$\lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) + \lambda\left\|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right\|_2 =$$

$$(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix} +$$
$$\lambda\mathcal{P}_j\left(\tilde{\boldsymbol{\alpha}}_j^g \circ \mathbb{1}_{\bar{S}}\right) + \lambda\mathcal{P}_j\left(\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) \circ \mathbb{1}_S\right) + \lambda\left\|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right\|_2 \leq$$

$$\frac{\lambda 2|S|^{1/2}}{C}\left\{(n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix}\right\}^{1/2} \leq$$

$$\frac{\lambda^2 4|S|}{C^2} + (n^g)^{-1}\begin{pmatrix}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right)^\top & \left(\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\right)^\top\end{pmatrix}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\end{pmatrix},$$

where we use the compatiblility condition for the first inequality, and for the second inequality use the fact that

$$ab \leq b^2 + a^2$$

for any $a, b \in \mathbb{R}$. The conclusion follows immediately. □

If the event $\mathcal{E}$ occurs with probability tending to one, Theorem B.2 implies

$$\mathcal{P}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) + \|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\|_2 = O_P\left(\lambda\right).$$

We select λ so that the event $\mathcal{E}$ occurs with high probability. For instance, suppose the elements of the matrix

$$\xi = \begin{pmatrix} \text{diag}\left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*}\right) \mathcal{V}_{j,1}^g + \mathcal{U}_{j,1}^g \\ \text{diag}\left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*}\right) \mathcal{V}_{j,2}^g + \mathcal{U}_{j,2}^g \end{pmatrix}$$

are sub-Gaussian, and consider the event

$$\bar{\mathcal{E}} = \left\{ \left\| \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \\ \left(\mathcal{V}_{j,2}^g\right)^\top \left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*}\right) + \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1} \end{pmatrix} \right\|_\infty \leq \frac{n^g \lambda_0}{d} \right\},$$

where $\|\cdot\|_\infty$ is the ℓ_∞ norm. Observing that $\mathcal{E} \subset \bar{\mathcal{E}}$, it is only necessary to show that $\bar{\mathcal{E}}$ holds with high probability. It is shown in Corollary 2 of Negahban et al. (2012) that there exist constants $u_1, u_2 > 0$ such that with $\lambda_0 \propto \{\log(p)/n\}^{1/2}$, $\bar{\mathcal{E}}$ holds with probability at least $1 - u_1 p^{-u_2}$. Thus, $\mathcal{E}$ occurs with probability tending to one as $p \to \infty$. For distributions with heavier tails, a larger choice of λ may be required (Yu et al., 2019).

B.2.4 De-biased Score Matching Estimator

The KKT conditions for the regularized score matching loss imply that the estimator $\tilde{\boldsymbol{\alpha}}_j^g$ satisfies

$$\nabla L_{n,j}(\tilde{\boldsymbol{\alpha}}_j^g, \tilde{\boldsymbol{\theta}}_j^g) = (n^g)^{-1} \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_j^g \\ \tilde{\boldsymbol{\theta}}_j^g \end{pmatrix} + (n^g)^{-1} \begin{pmatrix} \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \\ \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1} \end{pmatrix} = \begin{pmatrix} \lambda \nabla P\left(\tilde{\boldsymbol{\alpha}}_j^g\right) \\ \mathbf{0} \end{pmatrix}.$$

With some algebra, we can rewrite the KKT conditions as

$$(n^g)^{-1}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top\mathcal{V}_{j,2}^g\\ \left(\mathcal{V}_{j,2}^g\right)^\top\mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top\mathcal{V}_{j,2}^g\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g-\boldsymbol{\alpha}_j^{g,*}\\ \tilde{\boldsymbol{\theta}}_j^g-\boldsymbol{\theta}_j^{g,*}\end{pmatrix}=$$
$$\lambda\begin{pmatrix}\nabla P\left(\tilde{\boldsymbol{\alpha}}_j^g\right)\\ \mathbf{0}\end{pmatrix}-(n^g)^{-1}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,1}^g\right)^\top\mathbf{1}\\ \left(\mathcal{V}_{j,2}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,2}^g\right)^\top\mathbf{1}\end{pmatrix}.$$

Now, let $\Sigma_{j,n}$ be the matrix

$$\Sigma_{j,n}=(n^g)^{-1}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top\mathcal{V}_{j,2}^g\\ \left(\mathcal{V}_{j,2}^g\right)^\top\mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top\mathcal{V}_{j,2}^g\end{pmatrix},$$

let $\Sigma_j=E[\Sigma_{j,n}]$, and let $\tilde{M}_j$ be an estimate of Σ_j^{-1}. We can now rewrite the KKT conditions as

$$\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g-\boldsymbol{\alpha}_j^{g,*}\\ \tilde{\boldsymbol{\theta}}_j^g-\boldsymbol{\theta}_j^{g,*}\end{pmatrix}=\underbrace{\lambda\tilde{M}_j\begin{pmatrix}\nabla P\left(\tilde{\boldsymbol{\alpha}}_j^g\right)\\ \mathbf{0}\end{pmatrix}}_{\text{(i)}}-\underbrace{(n^g)^{-1}\,\tilde{M}_j\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,1}^g\right)^\top\mathbf{1}\\ \left(\mathcal{V}_{j,2}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,2}^g\right)^\top\mathbf{1}\end{pmatrix}}_{\text{(ii)}}+$$
$$\underbrace{(n^g)^{-1}\left\{I-\Sigma_{j,n}\tilde{M}_j\right\}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}_j^g-\boldsymbol{\alpha}_j^{g,*}\\ \tilde{\boldsymbol{\theta}}_j^g-\boldsymbol{\theta}_j^{g,*}\end{pmatrix}}_{\text{(iii)}}. \tag{B.4}$$

As is the case for the de-biased group LASSO in Appendix A, the first term (i) in (B.4) depends only on the observed data and can be directly subtracted from the initial estimate. The second term (ii) is asymptotically equivalent to

$$(n^g)^{-1}\,\Sigma_j^{-1}\begin{pmatrix}\left(\mathcal{V}_{j,1}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,1}^g\right)^\top\mathbf{1}\\ \left(\mathcal{V}_{j,2}^g\right)^\top\left(\mathcal{V}_{j,1}^g\boldsymbol{\alpha}_j^{g,*}+\mathcal{V}_{j,2}^g\boldsymbol{\theta}_j^{g,*}\right)+\left(\mathcal{U}_{j,2}^g\right)^\top\mathbf{1}\end{pmatrix}, \tag{B.5}$$

if $\tilde{M}_j$ is a consistent estimate of Σ_j^{-1}. Using the fact that $\mathbb{E}\left[\nabla L_{n,j}^g\left(\boldsymbol{\alpha}_j^{g,*},\boldsymbol{\theta}_j^{g,*}\right)\right]=\mathbf{0}$, it can be seen that (B.5) is an average of *i.i.d.* random quantities with mean zero. The central limit theorem then implies that any low-dimensional sub-vector is asymptotically normal. The last term (iii) is asymptotically negligible if $\tilde{M}_j$ is an approximate inverse of $\Sigma_{j,n}$ and

if $(\tilde{\boldsymbol{\alpha}}_j^g, \tilde{\boldsymbol{\theta}}_j^g)$ is consistent for $(\boldsymbol{\alpha}_j^{g,*}, \boldsymbol{\theta}_j^{g,*})$. Thus, for an appropriate choice of $\tilde{M}_j$, we expect asymptotic normality of an estimator of the form

$$\begin{pmatrix} \check{\boldsymbol{\alpha}}_j^g \\ \check{\boldsymbol{\theta}}_j^g \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_j^g \\ \tilde{\boldsymbol{\theta}}_j^g \end{pmatrix} - \lambda \tilde{M}_j \begin{pmatrix} \nabla P\left(\tilde{\boldsymbol{\alpha}}_j^g\right) \\ \mathbf{0} \end{pmatrix}.$$

Before constructing $\tilde{M}_j$, we first provide an alternative expression for Σ_j^{-1}. We define the $d \times d$ matrices $\Gamma_{j,k,l}^*$ and $\Delta_{j,k}^*$ as

$$\Gamma_{j,k,1}^*, \ldots, \Gamma_{j,k,p}^*, \Delta_{j,k}^* = \underset{\Gamma_1,\ldots,\Gamma_p,\Delta \in \mathbb{R}^{d\times d}}{\arg\min} \mathbb{E}\left[\operatorname{trace}\left\{(n^g)^{-1}\left(\mathcal{V}_{j,k,1}^g - \sum_{l\neq k,j} \mathcal{V}_{j,l,1}^g \Gamma_l - \mathcal{V}_{j,2}^g \Delta\right)^\top \left(\mathcal{V}_{j,k,1}^g - \sum_{l\neq k,j} \mathcal{V}_{j,l,1}^g \Gamma_l - \mathcal{V}_{j,2}^g \Delta\right)\right\}\right].$$

We also define the $d \times d$ matrices $\Lambda_{j,k}^*$ as

$$\Lambda_{j,1}^*, \ldots, \Lambda_{j,p}^* = \underset{\Lambda_1,\ldots,\Lambda_p,\in \mathbb{R}^{d\times d}}{\arg\min} E\left[\operatorname{trace}\left\{(n^g)^{-1}\left(\mathcal{V}_{j,2}^g - \sum_{k\neq j} \mathcal{V}_{j,k,1}^g \Lambda_k\right)^\top \left(\mathcal{V}_{j,2}^g - \sum_{k\neq j} \mathcal{V}_{j,k,1}^g \Lambda_k\right)\right\}\right].$$

Additionally, we define the $d \times d$ matrices $C_{j,k}^*$ and D_j^*

$$C_{j,k}^* = \mathbb{E}\left[(n^g)^{-1}\left(\mathcal{V}_{j,k,1}^g\right)^\top \left(\mathcal{V}_{j,k,1}^g - \sum_{l\neq k,j} \mathcal{V}_{j,l,1}^g \Gamma_{j,k,l}^* - \mathcal{V}_{j,2}^g \Delta_{j,k}^*\right)\right]$$
$$D_j^* = \mathbb{E}\left[(n^g)^{-1}\left(\mathcal{V}_{j,2}^g\right)^\top \left(\mathcal{V}_{j,2}^g - \sum_{k\neq j} \mathcal{V}_{j,k,1}^g \Lambda_{j,k}^*\right)\right].$$

It can be shown that Σ_j^{-1} can be expressed as

$$\Sigma_j^{-1} = \begin{pmatrix} \left(C_{j,1}^*\right)^{-1} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \left(C_{j,p}^*\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \left(D_j^*\right)^{-1} \end{pmatrix} \begin{pmatrix} I & -\Gamma_{j,1,2}^* & \cdots & -\Gamma_{j,1,p}^* & -\Delta_{j,1}^* \\ -\Gamma_{j,2,1}^* & I & \cdots & -\Gamma_{j,2,p}^* & -\Delta_{j,2}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\Gamma_{j,p,1}^* & -\Gamma_{j,p,2}^* & \cdots & I & -\Delta_{j,p}^* \\ -\Lambda_{j,1}^* & -\Lambda_{j,2}^* & \cdots & -\Lambda_{j,p}^* & I \end{pmatrix}.$$

We can thus estimate Σ_j^{-1} by estimating each of the matrices $\Gamma_{j,k,l}^*$, $\Lambda_{j,k}^*$, and $\Delta_{j,k}^*$.

Similar to our discussion of the de-biased group LASSO in Appendix A, we use a group-penalized variant of the nodewise LASSO to construct $\tilde{M}_j$. We estimate $\Gamma^*_{j,k,l}$ and $\Delta^*_{j,k}$ as

$$\tilde{\Gamma}_{j,k,1}, \ldots, \tilde{\Gamma}_{j,k,p}, \tilde{\Delta}_{j,k} =$$
$$\underset{\Gamma_1,\ldots,\Gamma_p,\Delta \in \mathbb{R}^{d\times d}}{\arg\min} \operatorname{trace}\left\{ (n^g)^{-1} \left(\mathcal{V}^g_{j,k,1} - \sum_{l\neq k,j} \mathcal{V}^g_{j,l,1}\Gamma_l - \mathcal{V}^g_{j,2}\Delta \right)^\top \left(\mathcal{V}^g_{j,k,1} - \sum_{l\neq k,j} \mathcal{V}^g_{j,l,1}\Gamma_l - \mathcal{V}^g_{j,2}\Delta \right) \right\} +$$
$$\omega_1 \sum_{l\neq k,j} \|\Gamma_l\|_{2,*} + \omega_1 \|\Delta\|_{2,*},$$

where $\omega_1, \omega_2 > 0$ are tuning parameters, and $\|\cdot\|_{2,*}$ is as defined in Appendix A. We estimate $\Lambda^*_{j,k}$ as

$$\tilde{\Lambda}_{j,1}, \ldots, \tilde{\Lambda}_{j,p} =$$
$$\underset{\Lambda_1,\ldots,\Lambda_p,\in \mathbb{R}^{d\times d}}{\arg\min} \operatorname{trace}\left\{ (n^g)^{-1} \left(\mathcal{V}^g_{j,2} - \sum_{k\neq j} \mathcal{V}^g_{j,k,1}\Lambda_k \right)^\top \left(\mathcal{V}^g_{j,2} - \sum_{k\neq j} \mathcal{V}^g_{j,k,1}\Lambda_k \right) \right\} + \omega_2 \sum_{l\neq j} \|\Lambda_k\|_{2,*}.$$

Additionally, we define the $d \times d$ matrices $\tilde{C}_{j,k}$ and $\tilde{D}_j$

$$\tilde{C}_{j,k} = (n^g)^{-1} \left(\mathcal{V}^g_{j,k,1}\right)^\top \left(\mathcal{V}^g_{j,k,1} - \sum_{l\neq k,j} \mathcal{V}^g_{j,l,1}\tilde{\Gamma}_{j,k,l} - \mathcal{V}^g_{j,2}\tilde{\Delta}_{j,k} \right)$$
$$\tilde{D}_j = (n^g)^{-1} \left(\mathcal{V}^g_{j,2}\right)^\top \left(\mathcal{V}^g_{j,2} - \sum_{k\neq j} \mathcal{V}^g_{j,k,1}\tilde{\Lambda}_{j,k} \right).$$

We then take $\tilde{M}_j$ as

$$\tilde{M}_j = \begin{pmatrix} \tilde{C}^{-1}_{j,1} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \tilde{C}^{-1}_{j,p} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \tilde{D}^{-1}_j \end{pmatrix} \begin{pmatrix} I & -\tilde{\Gamma}_{j,1,2} & \cdots & -\tilde{\Gamma}_{j,1,p} & -\tilde{\Delta}_{j,1} \\ -\tilde{\Gamma}_{j,2,1} & I & \cdots & -\tilde{\Gamma}_{j,2,p} & -\tilde{\Delta}_{j,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\tilde{\Gamma}_{j,p,1} & -\tilde{\Gamma}_{j,p,2} & \cdots & I & -\tilde{\Delta}_{j,p} \\ -\tilde{\Lambda}_{j,1} & -\tilde{\Lambda}_{j,2} & \cdots & -\tilde{\Lambda}_{j,p} & I \end{pmatrix}.$$

When $\Gamma^*_{j,k,l}$, $\Delta^*_{j,k}$, and $\Lambda^*_{j,k}$ satisfy appropriate sparsity conditions and some additional regularity assumptions, $\tilde{M}_j$ is a consistent estimate of Σ^{-1}_j for $\omega_1 \propto \{\log(p)/n\}^{1/2}$ and

$\omega_2 \propto \{\log(p)/n\}^{1/2}$ (see, e.g., Chapter 8 of Bühlmann and van de Geer, 2011) for a more comprehensive discussion). Using the same argument presented in Appendix A, we are able to obtain the following bound on a scaled version of the remainder term (iii):

$$\left\| \begin{pmatrix} \tilde{C}_{j,1} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \tilde{C}_{j,p} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \tilde{D}_j \end{pmatrix} \left\{ I - (n^g)^{-1} \begin{pmatrix} \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,1}^g\right)^\top \mathcal{V}_{j,2}^g \\ \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,1}^g & \left(\mathcal{V}_{j,2}^g\right)^\top \mathcal{V}_{j,2}^g \end{pmatrix} \tilde{M}_j \right\} \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*} \\ \tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*} \end{pmatrix} \right\|_\infty \leq$$

$$\max\{\omega_1, \omega_2\} \left\{ \mathcal{P}\left(\tilde{\boldsymbol{\alpha}}_j^g - \boldsymbol{\alpha}_j^{g,*}\right) + \|\tilde{\boldsymbol{\theta}}_j^g - \boldsymbol{\theta}_j^{g,*}\|_2 \right\}.$$

The remainder is $o_P(n^{-1/2})$ and hence asymptotically negligible if $n^{1/2} \max\{\omega_1, \omega_2\}\lambda \to 0$, where λ is the tuning parameter for the regularized score matching estimator (see Theorem B.2).

The de-biased estimate $\check{\alpha}_{j,k}^g$ of $\alpha_{j,k}^{g,*}$ can be expressed as

$$\check{\alpha}_{j,k}^g = \tilde{\alpha}_{j,k}^g - (n^g)^{-1} \tilde{C}_{j,k}^{-1} \left(\mathcal{V}_{j,k,1}^g - \sum_{l \neq j,k} \mathcal{V}_{j,l,1}^g \tilde{\Gamma}_{j,k,l} \right)^\top \left(\mathcal{V}_{j,1}^g \tilde{\boldsymbol{\alpha}}_j^g + \mathcal{V}_{j,2}^g \tilde{\boldsymbol{\theta}}_j^g + \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \right).$$

The difference between the de-biased estimator $\check{\alpha}_{j,k}^g$ and the true parameter $\alpha_{j,k}^{g,*}$ can be expressed as

$$\begin{aligned} \tilde{C}_{j,k}\left(\check{\alpha}_{j,k}^g - \alpha_{j,k}^{g,*}\right) = &- (n^g)^{-1} \left(\mathcal{V}_{j,k,1}^g - \sum_{l \neq j,k} \mathcal{V}_{j,l,1}^g \Gamma_{j,k,l}^* \right)^\top \left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*} + \left(\mathcal{U}_{j,1}^g\right)^\top \mathbf{1} \right) + \\ &(n^g)^{-1} \left(\mathcal{V}_{j,2}^g \Delta_{j,k}^* \right)^\top \left(\mathcal{V}_{j,1}^g \boldsymbol{\alpha}_j^{g,*} + \mathcal{V}_{j,2}^g \boldsymbol{\theta}_j^{g,*} + \left(\mathcal{U}_{j,2}^g\right)^\top \mathbf{1} \right) \Bigg\} + o_P\left(n^{-1/2}\right). \end{aligned}$$

As discussed above, the central limit theorem implies asymptotic normality of $\check{\alpha}_{j,k}^g$. We can estimate the asymptotic variance of $\check{\alpha}_{j,k}^g$ as

$$(n^g)^{-2} \tilde{C}_{j,k}^{-1} \tilde{M}_{j,k} \tilde{\xi}^\top \tilde{\xi} \tilde{M}_{j,k}^\top \left(\tilde{C}_{j,k}^{-1}\right)^\top,$$

where we define

$$\tilde{\xi} = \begin{pmatrix} \operatorname{diag}\left(\mathcal{V}_{j,1}^g \tilde{\boldsymbol{\alpha}}_j^g + \mathcal{V}_{j,2}^g \tilde{\boldsymbol{\theta}}_j^g \right) \mathcal{V}_{j,1}^g + \mathcal{U}_{j,1}^g \\ \operatorname{diag}\left(\mathcal{V}_{j,1}^g \tilde{\boldsymbol{\alpha}}_j^g + \mathcal{V}_{j,2}^g \tilde{\boldsymbol{\theta}}_j^g \right) \mathcal{V}_{j,2}^g + \mathcal{U}_{j,2}^g \end{pmatrix}$$

$$\tilde{M}_{j,k} = \begin{pmatrix} -\tilde{\Gamma}_{j,k,1} & \cdots & -\tilde{\Gamma}_{j,k,k-1} & I & -\tilde{\Gamma}_{j,k,k+1} & \cdots & -\tilde{\Gamma}_{j,k,p} & -\tilde{\Delta}_{j,p} \end{pmatrix}.$$

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Additional technical details

Estimation of Gâteaux derivative in Example 2

Recall the form of the Gâteaux derivative of the risk functional for the partially additive mean regression model in (4.5). The next result provides conditions under which the plug-in estimator in (4.6) is uniformly asymptotically linear with the efficient influence function defined in (4.7). Below, we denote by $\phi_{n,\theta_*}(\cdot;h)$ the estimated influence function

$$z=(w,x,y)\mapsto \phi_{n,\theta_*}(z;h):=\{y-\mu_{n,Y,P_0}(w)+\mu_{n,\theta_*,P_0}(w)-\theta_*(x)\}\{h(x)-\mu_{n,h,P_0}(w)\}-\dot{R}_{n,\theta_*}(h)\ .$$

Result C.1. *Suppose that there exists a P_0-Donsker class Φ such that $\phi_{P_0,\theta_*}(\cdot;h)$ and $\phi_{n,\theta_*}(\cdot;h)$ are both in Φ for all $h\in\mathcal{H}$ with probability tending to one. Furthermore, suppose that the rate conditions $\int\{\mu_{n,Y,P_0}(w)-\mu_{Y,P_0}(w)\}^2\,dP_0(w)=o_P(n^{-1/2})$, $\int\{\mu_{n,\theta_*,P_0}(w)-\mu_{\theta_*,P_0}(w)\}^2\,dP_0(w)=o_P(n^{-1/2})$, and*

$$\sup_{h\in\mathcal{H}}\int\{\mu_{n,h,P_0}(w)-\mu_{h,P_0}(w)\}^2\,dP_0(w)=o_P(n^{-1/2})$$

are satisfied. Then, it follows that $\sup_{h\in\mathcal{H}}|r_n(h)|=o_P(n^{-1/2})$, where $r_n(h)$ is the remainder term in (4.4).

Proof of Result C.1. The remainder term can be written as $r_n(h)=A_n(h)+B_n(h)$, where we define

$$A_n(h):=\int\{\phi_{n,\theta_*}(z;h)-\phi_{P_0,\theta_*}(z;h)\}\,d(P_n-P_0)(z)$$
$$B_n(h):=\int\{\phi_{n,\theta_*}(z;h)-\phi_{P_0,\theta_*}(z;h)\}\,dP_0(z)+\dot{R}_{n,\theta_*}(h)-\dot{R}_{0,\theta_*}(h).$$

For the first term, it is shown in the proof of Lemma 19.26 of van der Vaart (2000) that $\sup_{h\in\mathcal{H}} |A_n(h)| = o_P(n^{-1/2})$ in view of the Donsker class condition and uniform consistency of the nuisance parameter estimator. The second term can be expressed as

$$B_n(h) = \int \{\mu_{Y,P_0}(w) - \mu_{n,Y,P_0}(w) + \mu_{n,\theta_*,P_0}(w) - \mu_{\theta_*,P_0}(w)\} \{\mu_{n,h,P_0}(w) - \mu_{h,P_0}(w)\} \, dP_0(z).$$

By an application of the Cauchy-Schwarz inequality, and in view of the rate conditions, it follows that $\sup_{h\in\mathcal{H}} |B_n(h)| = o_P(n^{-1/2})$. This completes the proof in view of the triangle inequality. □

C.2 Proof of lemma and theorems

Proof of Lemma 4.1. The result follows immediately from Slutsky's theorem (see, e.g., Theorem 7.15 of Kosorok, 2008). □

Proof of Theorem 4.1. Below, denote by $\mathcal{B}$ the collection containing each bounded Lipschitz functionals $g : \ell^\infty(\mathcal{H}) \to [-1, 1]$ with Lipschitz constant 1, that is, satisfying that $|g(a_1) - g(a_2)| \leq \|a_1 - a_2\|_{\mathcal{H}}$ for all $a_1, a_2 \in \ell^\infty(\mathcal{H})$. Define the real-valued random functionals $\mathbb{G}_n : h \mapsto n^{-1/2} \sum_{i=1}^n \xi_i \phi_{f_n,\theta_*}(Z_i; h)$ and $\mathbb{G}_{n,0} : h \mapsto n^{-1/2} \sum_{i=1}^n \xi_i \phi_{f_0,\theta_*}(Z_i; h)$ defined on $\mathcal{H}$. We will show that

$$\sup_{g\in\mathcal{B}} |\mathbb{E}_\xi [g(\mathbb{G}_n)] - \mathbb{E}_0 [g(\mathbb{G})]| \tag{C.1}$$

converges to zero in outer probability, where $\mathbb{E}_\xi$ denotes expectation over the distribution of $\xi_1, \xi_2, \ldots, \xi_n$, which implies the desired result since convergence of the expectation of bounded Lipschitz functions of a stochastic process is equivalent to weak convergence in view of the Portmanteau lemma (see, e.g., Lemma 18.9 of van der Vaart, 2000).

First, by the triangle inequality, we note that

$$\sup_{g\in\mathcal{B}} |\mathbb{E}_\xi [g(\mathbb{G}_n)] - \mathbb{E}_0 [g(\mathbb{G})]| \;\leq\; \sup_{g\in\mathcal{B}} |\mathbb{E}_\xi [g(\mathbb{G}_{n,0})] - \mathbb{E}_0 [g(\mathbb{G})]| + \sup_{g\in\mathcal{B}} |\mathbb{E}_\xi [g(\mathbb{G}_n) - g(\mathbb{G}_{n,0})]| \, .$$

We define the function $\bar{g} : \ell^\infty(\mathcal{H}) \to \mathbb{R}$ pointwise as $\bar{g}(u) := \min(1, \|u\|_{\mathcal{H}})$. Since $|g(u_1) - g(u_2)| \leq \min(2, \|u_1 - u_2\|_{\mathcal{H}})$ for any $g \in \mathcal{B}$ and $u_1, u_2 \in \ell^\infty(\mathcal{H})$, we get that

$$\sup_{g \in \mathcal{B}} \left| \mathbb{E}_\xi \left[g(\mathbb{G}_n) - g(\mathbb{G}_{n,0}) \right] \right| \;\leq\; 2 \left| \mathbb{E}_\xi \left[\bar{g}(\mathbb{G}_n - \mathbb{G}_{n,0}) \right] \right| .$$

Defining for each $z \in \mathcal{Z}$ the random functionals $\varphi_n(z) : h \mapsto \phi_{f_n,\theta_*}(z;h) - \phi_{f_0,\theta_*}(z;h)$ and $\overline{\varphi}_n(z) := \varphi_n(z) - \int \varphi_n(z) dP_0(z)$, we use the triangle inequality again to establish that

$$\sup_{g \in \mathcal{B}} \left| \mathbb{E}_\xi \left[g(\mathbb{G}_n) \right] - \mathbb{E}_0 \left[g(\mathbb{G}) \right] \right| \;\leq\; \sup_{g \in \mathcal{B}} \left| \mathbb{E}_\xi \left[g(\mathbb{G}_{n,0}) \right] - \mathbb{E}_0 \left[g(\mathbb{G}) \right] \right| + 2(A_n + B_n) \; ,$$

where we have defined

$$\begin{aligned} A_n \;&:=\; \left| \mathbb{E}_\xi \left[\bar{g} \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \varphi_n(Z_i) \right) - \bar{g} \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \overline{\varphi}_n(Z_i) \right) \right] \right| \\ B_n \;&:=\; \left| \mathbb{E}_\xi \left[\bar{g} \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \overline{\varphi}_n(Z_i) \right) \right] \right| \; . \end{aligned}$$

The first summand above converges to zero in outer probability by Theorem 2.9.6 of van der Vaart and Wellner (1996). We find that $A_n \geq 0$ tends to zero in probability by observing that

$$\begin{aligned} A_n \;&\leq\; \sup_{g \in \mathcal{B}} \left| \mathbb{E}_\xi \left[g \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \varphi_n(Z_i) \right) - g \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \overline{\varphi}_n(Z_i) \right) \right] \right| \\ &\leq\; \sup_{h \in \mathcal{H}} \left| \int \varphi_n(z)(h) dP_0(z) \right| \mathbb{E}_\xi \left| n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \right| \\ &\leq\; \left[\sup_{h \in \mathcal{H}} \int \varphi_n(z)(h)^2 dP_0(z) \, \mathbb{E}_\xi \left(n^{-1} \textstyle\sum_{i=1}^n \xi_i^2 \right) \right]^{1/2} \;=\; \left[\sup_{h \in \mathcal{H}} \int \varphi_n(z)(h)^2 dP_0(z) \right]^{1/2} \;=\; o_P(1) \; , \end{aligned}$$

where the first inequality holds because $\bar{g}$ resides in $\mathcal{B}$, the second inequality holds in view of the Lipschitz property, the third inequality follows from the Cauchy-Schwarz inequality and the fact that $\xi_1, \xi_2, \ldots, \xi_n$ are independent and have mean zero, the first equality holds because $\xi_1, \xi_2, \ldots, \xi_n$ have unit second moment, and the last statement follows by assumption. Finally, we focus on the term $B_n \geq 0$. By the triangle inequality and the fact that $\bar{g} \in \mathcal{B}$, we have that $B_n \leq B_{1n} + B_{2n}$, where

$$\begin{aligned} B_{1n} \;&:=\; \sup_{g \in \mathcal{B}} \left| \mathbb{E}_\xi \left[g \left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \overline{\varphi}_n(Z_i) \right) \right] - \mathbb{E}_0 \left[g \left(n^{1/2} \textstyle\int \varphi_n(z) d(P_n - P_0)(z) \right) \right] \right| \\ B_{2n} \;&:=\; \left| \mathbb{E}_0 \left[\bar{g} \left(n^{1/2} \textstyle\int \varphi_n(z) d(P_n - P_0)(z) \right) \right] \right| \; . \end{aligned}$$

We define $\nu(z) : (h, f) \mapsto \phi_{f,\theta_*}(z; h) - \phi_{f_0,\theta_*}(z; h)$ and $\bar{\nu}(z) : (h, f) \mapsto \nu(z) - \int \nu(z) dP_0(z)$ as real-valued functionals over $\mathcal{H} \times \mathcal{F}$. Define $\mathcal{F}_\delta := \{f \in \mathcal{F} : \|f - f_0\|_{\mathcal{F}} < \delta\}$, and let $\mathcal{D}$ be the collection containing each bounded Lipschitz functional $q : \ell^\infty(\mathcal{H} \times \mathcal{F}_\delta) \to [-1, 1]$ with Lipschitz constant 1. For any $\delta > 0$, since $\phi_{f_n,\theta_*}(\cdot; h) - \phi_{f_0,\theta_*}(\cdot; h)$ is in Φ_δ for each $h \in \mathcal{H}$ with probability tending to one, we have that

$$B_{1n} \leq \sup_{q \in \mathcal{D}} \left| \mathbb{E}_\xi \left[q\left(n^{-1/2} \textstyle\sum_{i=1}^n \xi_i \overline{\nu}(Z_i) \right) \right] - \mathbb{E}_0 \left[q\left(n^{1/2} \textstyle\int \nu(z) d(P_n - P_0)(z) \right) \right] \right|$$

with probability tending to one. Since Φ_δ is P_0-Donsker for small enough $\delta > 0$, we find that this upper bound for B_{1n} tends to zero in outer probability in view of Theorem 2.9.6 of van der Vaart and Wellner (1996). Finally, we argue that B_{2n} tends to zero (deterministically). Defining the random variable $G_n := \bar{g}\left(n^{1/2} \int \varphi_n(z) d(P_n - P_0)(z)\right)$, we note that the sequence $\{G_1, G_2, \ldots\}$ is uniformly bounded by one. Furthermore, because $0 \leq G_n \leq \sup_{h \in \mathcal{H}} |n^{1/2} \int \varphi_n(z)(h) d(P_n - P_0)(z)|$, G_n tends to zero in probability provided

$$\sup_{h \in \mathcal{H}} \left| n^{1/2} \int \varphi_n(z)(h) d(P_n - P_0)(z) \right| = o_P(1) \ ,$$

in which case it follows that $B_{2n} = |\mathbb{E}_0(G_n)|$ tends to zero as well. This condition is shown in the proof of Lemma 19.26 of van der Vaart (2000), thereby completing the proof. □

Proof of Theorem 4.2. For conciseness, for any given probability P and P-integrable function f, we denote $\int f(z) dP(z)$ by the shorthand notation Pf below, and we write $\mathbb{G}_n := n^{1/2}(P_n - P_0)$. We begin by proving (a). By application of the continuous mapping theorem and Slutsky's theorem, the result follows if we can show that

$$\sup_{h \in \mathcal{H}_n} \mathbb{G}_n \phi_h = \sup_{h \in \mathcal{H}} \mathbb{G}_n \phi_h + o_P(1) \ . \tag{C.2}$$

By (4.10), there exists a deterministic sequence $\epsilon_n \downarrow 0$ such that

$$\mathbb{P}_0 \left(\sup_{h_1 \in \mathcal{H}} \inf_{h_2 \in \mathcal{H}_n} P_0(\phi_{h_1} - \phi_{h_2})^2 > \epsilon_n \right) \longrightarrow 0 \ .$$

By the definition of the supremum and infimum, for each n, there exist random functions $h_{1,n} \in \mathcal{H}$ and $h_{2,n} \in \mathcal{H}_n$ such that $\mathbb{G}_n \phi_{h_{1,n}} \geq \sup_{h \in \mathcal{H}} \mathbb{G}_n \phi_h - \epsilon_n$ and $P_0(\phi_{h_{1,n}} - \phi_{h_{2,n}})^2 \leq$

$\inf_{h\in\mathcal{H}_n} P_0(\phi_{h_{1,n}} - \phi_h)^2 + \epsilon_n$. Now, with probability tending to one,

$$\inf_{h\in\mathcal{H}} P_0(\phi_{h_{1,n}} - \phi_h)^2 \leq \sup_{h_1\in\mathcal{H}} \inf_{h_2\in\mathcal{H}_n} P_0(\phi_{h_1} - \phi_{h_2})^2 \leq \epsilon_n \ ,$$

and so, $P_0(\phi_{h_{1,n}} - \phi_{h_{2,n}})^2 = o_P(1)$. We can write $\sup_{h\in\mathcal{H}} \mathbb{G}_n\phi_h - \sup_{h\in\mathcal{H}_n} \mathbb{G}_n\phi_h = A_n + B_n + C_n$ with $A_n := \sup_{h\in\mathcal{H}} \mathbb{G}_n(\phi_h - \phi_{h_{1,n}})$, $B_n := \inf_{h\in\mathcal{H}_n} \mathbb{G}_n(\phi_{h_{2,n}} - \phi_h)$ and $C_n := \mathbb{G}_n(\phi_{h_{1,n}} - \phi_{h_{2,n}})$. By construction, we have that $0 \leq A_n \leq \epsilon_n$, and so, $A_n = o_P(1)$. Additionally, we have that $B_n \leq 0$ because $h_{2,n} \in \mathcal{H}_n$. Finally, since $P_0(\phi_{h_{1,n}} - \phi_{h_{2,n}})^2 = o_P(1)$ and $\phi_{h_{1,n}} - \phi_{h_{2,n}}$ belongs to the Donsker class $\{\phi_{h_1} - \phi_{h_2} : h_1, h_2 \in \bar{\mathcal{H}}\}$, we have that $C_n = o_P(1)$ by application of Lemma 19.24 of van der Vaart (2000). Thus, we have established that $\sup_{h\in\mathcal{H}} \mathbb{G}_n\phi_h - \sup_{h\in\mathcal{H}_n} \mathbb{G}_n\phi_h$ is bounded above by an $o_P(1)$ term. A similar argument can be used to conclude the same of $\sup_{h\in\mathcal{H}_n} \mathbb{G}_n\phi_h - \sup_{h\in\mathcal{H}} \mathbb{G}_n\phi_h$, thus establishing (C.2).

We now prove part (b). We note that

$$\left| \int_{\mathcal{H}_n} [(P_n - P_0)\phi_h]^2 d\bar{Q}(h) - \int_{\mathcal{H}} [(P_n - P_0)\phi_h]^2 d\bar{Q}(h) \right| \leq$$
$$\int_{(\mathcal{H}_n\cup\mathcal{H})\setminus(\mathcal{H}_n\cap\mathcal{H})} [(P_n - P_0)\phi_h]^2 d\bar{Q}(h) \leq$$
$$\bar{Q}\left(\{\mathcal{H}\cup\mathcal{H}_n\}\setminus\{\mathcal{H}\cap\mathcal{H}_n\}\right) \left[\sup_{h\in\bar{\mathcal{H}}} |(P_n - P_0)\phi_h|\right]^2 .$$

Since $\bar{\mathcal{H}}$ is a Donsker class and (4.11) holds by assumption, this implies that

$$\int_{\mathcal{H}_n} [(P_n - P_0)\phi_h]^2 d\bar{Q}(h) = \int_{\mathcal{H}} [(P_n - P_0)\phi_h]^2 d\bar{Q}(h) + o_P(n^{-1}) \ ,$$

and the result follows by an application of Slutsky's theorem and the continuous mapping theorem. □