©Copyright 2021
Yunhua Xiang

Methods and Theory for Nonparametric Inference In High-dimensional Settings

Yunhua Xiang

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:
Noah Simon, Chair
Ali Shojaie
Michael Wu

Program Authorized to Offer Degree:
Biostatistics-Public Health

University of Washington

Abstract

Methods and Theory for Nonparametric Inference In High-dimensional Settings

Yunhua Xiang

Chair of the Supervisory Committee:
Dr. Noah Simon
Department of Biostatistics

This dissertation addresses nonparametric estimation and inference problems of graphical modeling, linear association assessment, and matrix completion. First, we introduce a flexible framework for nonparametric graphical modeling. We propose three nonparametric measures of conditional dependence, which have theoretically optimal estimators that allow incorporation of flexible machine learning techniques and yield wald-type confidence intervals. In the second project, we propose a nonparametric parameter to measure the linear association between the outcome and explanatory variables. This parameter is always explicitly defined even when the true relationship is nonlinear and is equivalent with the regression coefficient under a linear model space. Thus, its estimator can be a more robust alternative to the standard model-based techniques to estimate the coefficients of a linear model. In the final project, we theoretically show that nuclear-norm penalization used for recovering low-rank matrices, remains effective even when the underlying matrices are generated by a low-dimensional non-linear manifold. The convergence rate can be expressed as a function of the size of the matrix, as well as the smoothness and dimension of the manifold, which is minimax optimal (up to a log term).

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Time flies. I cannot believe that I am graduating and stepping into the next stage of my life. This journey, my twenty-year school life, is full of adventure and challenge, and would not be completed without the help of many people. Here, I want to express my sincere appreciation to them.

First, I would like to thank my dissertation advisor, Noah Simon, who gave me tremendous support, academic and non-academic. When I was a master student in Statistics, I took his class BIOST527, where he taught me a lesson that those complicated machine learning techniques could be explained so intuitively. When I decided to apply for a PhD program, he gave me school suggestions and wrote me a reference letter. After I came to the Department of Biostatistic, he became my advisor who helped me a lot in preparing the applied qualification, improving writing/presenting skills, as well as applying for a full-time job in industry. Most importantly, for my PhD research, he helped to figure out my dissertation projects, prove those technical theorems whenever I had difficulties, and edit manuscripts. I am really thankful to him for investing so much time in me and for encouraging me when I am not confident. I am really lucky to work with Noah and have him as my advisor throughout my time in UW.

I thank Ali Shojaie who also advised me on my third project, gave feedback on my manuscripts and served in my dissertation committee. Once upon a time, when I finished my first presentation in Slablab, he gave me a thump up, which I know was just an encouragement. But I was still very touched because I was so uncertain at that time. I would like to thank Mike Wu, who is my RA supervisor at Fred Hutch. I really appreciate that he never gave me pressure and always said kind words when he felt my anxiety. Every time when we met, I

felt like I was talking to a friend. He also gave me chances working with junior students and trusted me that I could help. I am very grateful that I could work with him. I also thank the rest of my doctoral committee members, Eardi Lila and Jennifer Balkus, for for their willingness to serve in my committee.

I thank Gitana Garofalo for her warm welcome when I joined the program and unwavering support to our students. She is always kind and shows a very positive way of treating people. She tries her best to make sure every student is supported and has a place to look for help when they are in need. I also want to thank Mauricio Sadinle and Lurdes Inoue, who have been my faculty advisor. When I was a junior student, they gave me suggestions about my progress and helped me with some administrative work. Thank you as well to my cohort and other students in the department. We learned statistics together, traveled together, and shared our stress together. You really make my PhD life more wonderful.

I also want to thank my friends from high school, undergrad and Amazon. Thank you, Huiwan Zhang, for being my friend for thirteen years. Although we currently live in different countries and cannot meet frequently. We still chat about everything by typing on Wechat. Thank you, Qiner Shi, for deciding to come to the US with me at our 20+ years of age. You let me know that I was not alone and now I have a place to stay in SF. Thank you, Ruiting Chen and Yinhong Lin. As I always said, the best thing I got from Amazon is not the working experience but is that I made friends with you and then later met so many other kind people. I cannot image how boring my PhD life would be under quarantine without you.

Finally, I want to thank my mom and dad for being so supportive and giving me a lot of love. No matter what big decisions I make, they always stand behind me and let me do what I want. When I feel upset and stressful, they just tell me that all they want is my happiness. Without them, I would not be who I am. As I know, there is at least one corner in the world, I am being loved. I am incredibly lucky to be your daughter and I love you too.

DEDICATION

to my parents, Ping Xiang and Zhilan Duan

Chapter 1

INTRODUCTION

In this chapter, we introduce the background and motivation of the three problems studied in this dissertation: graphical modeling, linear association assessment, and matrix completion. Many popular methods in these fields depend on some parametric specification of the data generating mechanism. This greatly impacts their usage in practice, e.g., resulting in bias of our estimators, and incorrect confidence interval coverage, as those parametric assumptions can be easily violated. The main contribution of this dissertation is the development of nonparametric methods for each of these problems, which allow us to leverage flexible machine learning techniques and accommodate for high-dimensional data.

1.1 Undirected Graphical Modeling

With the development of high-throughput measurement technologies in biotechnology, engineering, and elsewhere, it is increasingly common to measure a number of features without a strong apriori understanding on the interplay between them. It is fundamental to the development of science to learn these relationships. For example, it is important for biology to understand co-expression of genes [101]. Graphical modeling has been broadly useful to encode such a relationship, where two nodes/features are connected by an edge if they are dependent (according to some specified notion of dependence) while the weights of edges indicate the degree of dependence. Covariance and correlation are common dependence measures used in the network analysis[103, 130].

However, features can easily be connected due to *indirect effects* [9] using these measures. Instead, one is often interested in a more causally-motivated parameter: *conditional dependence*. That is, for two features Y and Z, we aim to assess if there an association between

them, while fixing all other features X. Gaussian Graphical Models (GGM) [129, 42, 109] have been developed to address this issue: features with non-zero entries in the precision matrix are connected. This precisely encodes conditional dependence when all of the features considered have a joint Gaussian distribution. But this is rarely the case in practice, and thus the detected edges may correspond to scientific quantities of little interest. This idea was extended by [71] and [7] to transelliptical graphical models, however the scientific relevance of the quantities estimated by these methods require similar (though slightly relaxed) assumptions.

There are other frameworks proposed to evaluate the degree of conditional dependence. [35] measured the local dependence of pairs via a conditional covariance function by monotonically transforming the conditioning function to a total score. [9] weakened the concept of conditional independence and applied the conditional correlation coefficient to account for the dependence structure. [43, 102] and [44] considered a more general nonparametric characterization of conditional independence using covariance operators on reproducing kernel Hilbert spaces (RKHS) to capture nonlinear dependence. However, in these cases, a *local parameter* was used: the conditional dependence measures depend on the value taken by conditioning variables. This parameter thus cannot be used as a summary measure.

In Chapter 2, we consider a more general form of conditional dependence: The dependence measures proposed in that chapter are nonparametric and summarize the average degree of association between a pair of features. Therefore, they can be useful for summarizing dependence without relying on strong parametric assumptions, and provide a single summary of dependence between features, (as compared to local quantities such as $\mathrm{Cov}(Y, Z|X)$). In addition, they admit simple and natural estimators, that facilitate the use of general machine learning methods, and allow us to construct asymptotically valid confidence intervals.

1.2 Conditionally Linear Association Measure

One of the most canonical problems in biostatistics is to understand the relationship between a response variable $Y \in \mathbb{R}$ and one or more explanatory variables $\mathbf{X} \in \mathbb{R}^p$. The most popular

tool is *linear regression*, which assumes the following linear relationship:

$$\mathrm{E}(Y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + \epsilon = X_j\beta_j + \mathbf{X}_{-j}\boldsymbol{\beta}_{-j}. \tag{1.1}$$

Under this model, it is clear that β_j plays the key role in assessing the degree of linear association between Y and X_j: The primary goal is to estimate those unknown coefficients $\beta = (\beta_1, ..., \beta_p)$. Ordinary least squares (OLS) estimation is standard technique when the dimension p is fixed and $p < n$. When the parametric (linear) model is correctly specified, the OLS estimator has favorable theoretical guarantees [76, 94, 120]. To address estimation of β when $p > n$, penalization techniques have been developed, such as the LASSO [104], which adds an l_1 penalty to the loss function and subsequently produces sparse estimators. While this penalty facilitates estimation of regression coefficients, it also biases the estimates resulting in problems when trying to characterize their limiting distributions. To combat this, [131] and [111] propose a debiased version of the lasso, that yields a non-sparse estimator. Under certain regularity conditions and sparsity assumptions, this debiased lasso estimator is asymptotically consistent and normal, and thus can be used to construct confidence intervals for β.

The validity of OLS and debiased lasso estimation is based on correct specification of a parametric (linear) model, i.e. X and Y must truly have a linear relationship. However, once this assumption is violated, those asymptotic properties do not hold anymore and the estimated β coefficients lose interpretability. In Chapter 3, we instead consider a nonparametric measure for quantifying the *linear association between* X *and* Y. This measure (i) does not require any parametric assumptions and can be more meaningful in assessing the degree of linear association between two variables (accounting for potential confounding by other features) when the linear model is misspecified; (ii) faciliates estimation with both low- and high-dimensional data; (iii) has a simple estimator that accommodates machine learning based models and a well characterized limiting distribution. More importantly, we show that this nonparametric parameter reduces to β under the presence of linearity, and, in that case,

the proposed estimator has the same limiting distribution as OLS estimator when $p < n$ and as the debiased lasso estimator when $p \gg n$.

1.3 Nuclear-norm-based Matrix Completion

Matrix completion is a framework that has gained popularity in a wide range of machine learning applications, including recommendation systems [63], system identification [72], and natural language processing [122]. It is a useful framework for complex prediction problems, where each observation comes with a heterogeneous collection of observed features. In particular, matrix completion is applied to problems where the object of inference or prediction is a matrix whose rows correspond to observations and columns to variables/features. In many cases, only a subset of entries in this matrix are observed (often with noise), and the goal is to "complete" the matrix, filling in estimates of the unobserved entries. The most famous example is the Netflix Challenge [63], where a small sample of observed ratings for each customer was used to successfully predict movie ratings for Netflix customers.

Assume we have an underlying unobserved matrix $M \in \mathbb{R}^{n \times p}$, and we only observe a subset of observations from the contaminated matrix $Y = M + E$, where E is a matrix of mean zero and finite variance noise variables. It is of interest to recover M from these noisy observations. However, without any structure in the partially observed matrix, filling in the unobserved entries in a meaningful way is impossible [66]. Matrix completion becomes possible if one imposes some constraints on the structure of the underlying matrix, and in particular, a constraint on the rank of the underlying matrix. Over the last decade, computationally efficient methods using convex optimization have been developed for recovering a low-rank matrix from a small number of observations with near-optimal statistical properties [99, 84, 21, 85, 61]. In particular, these methods rely on using the nuclear norm of the matrix [40]. The low-rank structure leveraged in matrix completion can be thought of as a linear embedding of the data in a low-dimensional space.

In practice, there is no reason to assume low-dimensional linear structure in the underlying matrix (as would be imposed by rank constraints). However the matrix may still have useful

low-dimensional structure. For example, low-dimensional nonlinear embeddings of data have been used fruitfully in motion recovery [124], epigenomics [93], and health data analytics [118]. To recover these embeddings, both Reproducing Kernel Hilbert Space (RKHS) methods [38], and deep neural networks [37] have been utilized.

Though initially designed for low-rank matrices, matrix completion methods have nevertheless been seen to perform well in cases where it seems unlikely that the underlying matrix has low rank [87, 36]. Despite many insights and algorithms, the success of matrix completion in these settings is not fully explained by existing theoretical results. In Chapter 4, we examine why matrix completion performs well in these scenarios. In particular, we show that if (i) the underlying true matrix to be recovered can be embedded in a low-dimensional manifold, and (ii) the curvature of that manifold is not too extreme, then nuclear-norm based matrix completion will consistently estimate the true underlying matrix. We additionally bound our reconstruction error as a function of the size of the matrix, the number of observed entries, the dimension of our embedding, and the curvature of our manifold. We also provide a lower bound, which shows that the nuclear-norm based estimator is minimax rate optimal (up to a log term).

Chapter 2

A FLEXIBLE FRAMEWORK FOR NONPARAMETRIC GRAPHICAL MODELING THAT ACCOMMODATES MACHINE LEARNING

2.1 Introduction

With the development of new high-throughput measurement technologies in biotechnology, engineering, and elsewhere, it is increasingly common to measure a number of features on each of a collection of people/objects without a strong apriori understanding on the interplay between these features. It is fundamental to developing science that we learn these relationships. For example, understanding co-expression of genes [101, 10, 73, 28] is foundational to biology; identifying regulatory networks [48] can help us understand cell differentiation [53, 15], and identify targets for treatment of disease [30, 12]; and among many other applications.

The relationships between features can be evaluated and expressed using *Graphical Modeling*: Here we use a graph $G = (V, E, W)$, where $V = \{1, ..., p\}$ $(p > 2)$ indexes a set of nodes $\{V_i\}_{i \in V}$ representing the features, $E = \{e_{i,j}\}$ is a set of edges corresponding to dependence between adjacent nodes, and $W = \{w_{i,j}\}$ is a collection of weights expressing the strength of each edge. In defining these edges and weights, one must decide on a measure of association/dependence. Covariance and correlation are two commonly-used measures for the dependence between two variables in multivariate analysis [2, 103, 89, 65, 27, 130].

However, one is often interested in a more causally-motivated parameter: In particular, when using correlation, features can easily be connected due to *indirect effects* [9]. For example, two "connected" features may be mechanistically tied to a third feature, and otherwise completely unrelated. These are often not the edges we wish to discover. One

is often more interested in a *conditional measure*: For two features Y and Z, conditional on fixing all other features $X = \{V_k\}_{k=1}^p - \{Y, Z\}$, we aim to assess if there an association between Y and Z. Previous work has attempted to address this using partial correlation [32, 4]. Rather than connecting features with non-zero correlation, instead features with non-zero entries in the precision matrix are connected. This corresponds to assessing the conditional dependence when all of the features considered have a joint Gaussian distribution [129, 42]. In practice, that is rarely, if ever, the case, and edges may correspond to scientific quantities of little interest.

In this chapter, we address this issue: We consider a more general form of conditional dependence that reduces to the partial correlation when all features are Gaussian. This dependence measure admits a straightforward, natural, and efficient estimator, that facilitates the use of general machine learning methods in estimating dependence. In addition, these estimators allow us to construct asymptotically-valid confidence intervals and run hypothesis tests (while accounting for multiple testing, when evaluating all edges in a graph).

The dependence measure that we primarily consider, which we term the *scaled expected conditional covariance* is

$$\Psi_{Y,Z} = \frac{\mathrm{E}\left[\mathrm{Cov}\left(Y, Z|X\right)\right]}{\sqrt{\mathrm{E}[\mathrm{Var}\left(Y|X\right)]}\sqrt{\mathrm{E}[\mathrm{Var}\left(Z|X\right)]}}. \tag{2.1}$$

Here, $\mathrm{Cov}(Y, Z|X)$ is the conditional covariance of Y and Z given X, and $\mathrm{Var}(Y|X)$ is the conditional variance of Y given X. This *parameter* is just a functional that maps the joint distribution of X, Y, and Z to a real number. In contrast to parameters from classical statistics, e.g. coefficients in a linear model, $\Psi_{Y,Z}$ is model agnostic, and does not implicitly assume any functional form on the relationships between our variables. This parameter summarizes the average degree of association between our features: This summarization using the *average* has two advantageous attributes: 1) It provides a *single* summary of dependence between features; and 2) Averages can be estimated at better rates than local quantities [14]. These issues dissuade us from directly using a local quantity such as $\mathrm{Cov}(Y, Z|X)$.

Later, we will further show that estimating these average dependence measures, such as (2.1), primarily (and in some cases only) relies on the estimation of a conditional mean. This reduces the problem of testing/evaluating conditional dependence to a canonical prediction problem, which allows us to naturally incorporate flexible machine learning techniques, such as generalized additive models [49], local polynomial regression [95], random forests [69] etc., and make inference even when X is high-dimensional [104, 75].

2.2 Related Work

Related work falls in two categories: The first does not directly estimate a parameter encoding dependence, but rather just tests a null hypothesis of conditional independence. This is the strategy generally taken with Gaussian graphical models [121, 105, 109], where the graph structure is encoded by the precision matrix. This idea was extended by [71] and [7] to transelliptical graphical model where nonparametric rank-based regularization estimators were used for estimating the latent inverse covariance matrix. Although, these approaches generalize the estimation to non-Gaussian setting and accommodate for high-dimensional data. They still assume specific underlying model structures.

The other approach evaluates the degree of dependence through estimation of a *target parameter*: [35] measured the local dependence of pairs via a conditional covariance function by monotonically transforming the conditioning function to a total score. [9] weakened the concept of conditional independence and applied the conditional correlation coefficient to account for the dependence structure. Others [43, 46, 102, 44] consider a more general nonparametric characterization of conditional independence using covariance operators on reproducing kernel Hilbert spaces (RKHS) to capture nonlinear dependence. However, in these cases, a *local parameter* was used: These conditional dependence measures depend on the value taken by conditioning variables. This parameter thus cannot be used as a summary measure.

Summary measures of conditional dependence which i) do not make parametric assumptions on the model; and ii) adjust for other covariates have been proposed in re-

gression setting. The most canonical of such measures is the average treatment effect $\int \mathrm{E}[Y|X=x,Z=1]-\mathrm{E}[Y|X=x,Z=0]dP(x)$ [8], which has been extensively discussed in the semiparametric context [114, 58]. But this measure is limited to evaluating association with a binary treatment. Approaches that attempt to use this with a continuous treatment are often either adhoc, or result in a *local measure* [51, 50, 59].

There exist methodologies which give omnibus measures of departure from conditional independence. For example, [132] and [117] used kernels and characteristic functions respectively to average over some functions of the conditioning variables. These methods have the potential advantage that they use an omnibus test and thus do not have to prespecify a particular direction to consider for departures from conditional independence. This advantage however is tied to their restriction: they need to specify very specific methods of "regressing out the conditioning variables", such as using RKHS regression or local averaging. This may be inappropriate when confounders are high-dimensional or with heterogeneous types. In addition, tuning of hyperparameters in these methods can be difficult. The theoretically optimal bandwidth pointed out in the paper can be hardly achievable by any sort of split sample validation criterion, such as minimizing MSE.

There are other methods which use resampling strategies to modify the original data, in an attempt to construct a pseudo-dataset where the indicated features are conditionally independent, [34] cleverly uses a restricted set of permutations that fix something akin to a sufficient dimension reduction of the conditioning variables. This approach works well in some scenarios, however with high-dimensional features, for example, it may be infeasible to effectively select such a dimension reduction, which would result in a procedure more akin to a marginal, rather than conditional independence testing. [96] uses a bootstrap to construct pseudo-conditionally-independent data. It then attempts to differentiate between the original data, and this new pseudo-data. Failure to differentiate suggests that the original data was conditionally independent. This methodology does allow ML-based tools to be used in constructing the classifier, however it still hinges on our ability to construct conditionally independent pseudo-data.

[79] recently discussed expected conditional covariance (one of the 3 measures in this chapter) as a summary of dependence in low-dimensional partially-linear additive regression. Their estimator is similarly a plug-in, however they discuss only a very particular strategy (which does not leverage Machine Learning techniques) of estimating the requisite conditional mean functions. In contrast, we decouple estimation of the conditional mean from evaluation of the expected conditional covariance. As such, in Section 2.4, we show that a wide array of ML-based predictive modeling techniques might be used in building those predictive functions for the conditional mean, and then leveraged in estimation of the expected conditional covariance.

2.3 Average Conditional Dependence Measures

Let $O = (Y, Z, X) \in \mathbb{R}^p$ denote a random vector drawn from some joint distribution $P \in \mathcal{M}$, where $\mathcal{M}$ is an unrestricted model space. Here, we have $Y \in \mathbb{R}, Z \in \mathbb{R}$, and $X \in \mathbb{R}^{p-2}$. For ease of notation, we have identified Y and Z as a pair of features of interest, and are aiming to evaluate the dependence between Y and Z conditional on X. However, we eventually plan to evaluate this dependence between all pairs of variables.

For simplicity, we denote the conditional means and the conditional variances with respect to distribution P as $\mu_{P,Y}(x) = \mathrm{E}_P(Y|X = x)$ and $\sigma^2_{P,Y}(x) = \mathrm{Var}_P(Y|X = x)$. Our first measure of dependence, previously mentioned in Section 2.1, is the *expected conditional covariance*

$$\begin{aligned} \Psi_1(P) &= \mathrm{E}_P[\mathrm{Cov}_P(Y, Z|X)] \\ &= \int (y - \mu_{P,Y}(x))(z - \mu_{P,Z}(x)) dP(o), \end{aligned} \tag{2.2}$$

We define our second measure similarly, as the *expected conditional correlation*

$$\begin{aligned} \Psi_2(P) &= \mathrm{E}_P[\mathrm{Corr}_P(Y, Z|X)] \\ &= \int \frac{\mathrm{Cov}_P(Y, Z|X = x)}{\sqrt{\sigma^2_{P,Y}(x)\sigma^2_{P,Z}(x)}} dP(x). \end{aligned} \tag{2.3}$$

Ψ_1 and Ψ_2 are the averaged conditional analogs to covariance and correlation. By averaging these conditional associations, these measures provide a global, instead of local, assessment of dependence.

In graphical modeling, as we are evaluating dependence between multiple pairs of features, it is important to use a standardized measure of association. Non-zero values of Ψ_1 will vary according to the scale of our variables. In contrast, Ψ_2 is standardized. Unfortunately, while Ψ_2 appears to be a very natural quantity, it ends up being somewhat difficult to estimate (this is further discussed in Section 2.4). In light of this, we propose a third, alternative standardized measure of dependence which we term the *scaled expected conditional covariance*

$$\Psi_3(P) = \frac{\Psi_1(P)}{\sqrt{V_Y(P)V_Z(P)}}, \tag{2.4}$$

where $V_Y(P) = \mathrm{E}_P[\sigma^2_{P,Y}(X)]$ and $V_Z(P) = \mathrm{E}_P[\sigma^2_{P,Z}(X)]$. Ψ_3 is constructed by scaling the expected conditional covariance with the square root of the products of the two expected conditional variances. This is analogous to how correlation is formed from covariance (only, in this case we average before taking our quotient). Indeed, it is simple to show that Ψ_3 is scale invariant, and furthermore takes on values in $[-1, 1]$.

Though Ψ_3 is perhaps less natural than Ψ_2, it turns out to be much easier to estimate from data. This makes intuitive sense as Ψ_2 contains positive *local* quantities in the denominator (the conditional standard deviations), where Ψ_3 contains only *global* quantities in the denominator. Estimating local quantities is more difficult, and instability of those estimates in the denominator (in particular if they are near 0) will result in instability of the estimator of Ψ_2. More specifically, our theory takes advantage of the fact that $V_Y(P) = \mathrm{E}_P[\mathrm{Cov}_P(Y, Y|X)]$, and that the standard delta-method can be applied to a ratio of efficient estimators in the case of Ψ_3 [82].

2.3.1 Higher Order Dependence

In this Section we discuss the relationship of our parameters to the conditional dependence/independence of features. In particular, we know that, without modification, covariance only encodes linear dependence. Unless variables are jointly Gaussian, linear independence does not imply independence [54]. However, general dependence can be evaluated using higher-order moments (or equivalently covariance of derived features) [44, 46]. Using similar ideas, we relate our dependence measures to non-linear association.

Consider two pre-specified functions $\phi_1 : \mathbb{R} \to \mathbb{R}^{p_1}$ and $\phi_2 : \mathbb{R} \to \mathbb{R}^{p_2}$ and assume that both functions are conditionally integrable: $\mathrm{E}[\phi_1(Y)|X] < \infty$, $\mathrm{E}[\phi_2(Z)|X] < \infty$. Further consider a non-negative weight function $w(x)$. Then, the (ϕ_1, ϕ_2, w)-expected conditional covariance is defined as

$$\Psi_1^{\phi_1,\phi_2,w}(P) = \mathrm{E}_P[w(x)\,\mathrm{Cov}_P(\phi_1(Y), \phi_2(Z)|X)]. \tag{2.5}$$

One can similarly extend $\Psi_2(P)$ and $\Psi_3(P)$ by replacing Y and Z with $\phi_1(Y)$ and $\phi_2(Z)$. Theoretically, estimating $\Psi_1^{\phi_1,\phi_2,w}(P)$ is essentially the same as estimating $\Psi_1(P)$ since $\phi_1(Y)$ is nothing more than a random variable. But conceptually, this simple transformation in (2.5) allows us assess higher order conditional dependence structure between Y and Z. In many cases, $w(x)$ will be taken to be 1, however it is required to characterize necessary and sufficient conditions for conditional independence.

2.3.2 Conditional Independence Testing

Using this idea of higher order dependence, we can develop necessary and sufficient conditions for conditional independence between Y and Z conditional on X. In particular, We consider (ϕ_1, ϕ_2, w)-expected conditional covariance, for $w(X) = \mathbb{1}\{X \in S_x\}$, $\phi_1(Y) = \mathbb{1}\{Y \in S_y\}$, and $\phi_2(Z) = \mathbb{1}\{Z \in S_z\}$ for arbitrary sets S_x, S_y, and S_z. In this case, we see that (ϕ_1, ϕ_2, w)-expected conditional covariance equal to 0 is equivalent to $P(Y \in S_y, Z \in S_z|X \in S_x) =$

$P(Y \in S_y | X \in S_x)P(Z \in S_z | X \in S_x)$. This gives us a simple necessary and sufficient condition for conditional independence

Proposition 2.1. *Random variables Y and Z are independent conditional on X iff for every ϕ_1, ϕ_2 and w in $\ell_2(P)$[1], for which $\Psi_1^{\phi_1,\phi_2,w}(P)$ is defined and finite, we have $\Psi_1^{\phi_1,\phi_2,w}(P) = 0$.*

Comprehensively testing for conditional independence via Proposition 2.1 is generally intractable as one would have to consider all possible w, ϕ_1, and ϕ_2. This is unsurprising: General conditional dependence is extremely difficult to evaluate — in practice impossible with any reasonable quantity of data in moderate to high-dimensions. In practice, we instead choose a few test functions (ϕ_1 and ϕ_2) to use, and just evaluate conditional dependence in those directions (finding conditional associations in any of those directions does imply that our features are *not* conditionally independent). This same idea is employed with Gaussian graphical modeling; only there, conditional dependence is completely characterized by linear conditional dependence. Additionally, in the joint Gaussian setting local and global dependence are equivalent (the conditional covariance between two features in a joint Gaussian model cannot vary with the values of the other features).

In the rest of this chapter, we just consider $\phi_1(y) = y$, $\phi_2(z) = z$, and $w(x) = 1$, returning to our original measures. While these measures cannot conclusively show that a pair of features are conditionally independent, if any of Ψ_1, Ψ_2 or Ψ_3 are non-zero, that does allow us to conclude that those features are conditionally dependent.

2.4 Estimating the Parameters

Suppose that we observe n i.i.d samples $\{o_i\}_{i=1}^n = \{y_i, z_i, x_i\}_{i=1}^n$ from an unknown distribution $P \in \mathcal{M}$ where $\mathcal{M}$ is a nonparametric model space. Our goal is to estimate the three well-defined global measures $\Psi_i, i = 1, 2, 3$ for conditional dependence. Before we discuss specific estimation of these 3 measures, we note that all 3 will require estimation of the intermediate quantities

[1] $\ell_2(P)$ represents a function class, where any function f in this class is square-integrable and measurable with respect to P.

$\mu_{P,Y}(x) = \mathrm{E}_P[Y|X]$ and $\mu_{P,Z}(x) = \mathrm{E}_P[Z|X]$. Estimating these conditional means is precisely the goal of most predictive modeling techniques. In the case that Y or Z is continuous, regression techniques can be used; If they are binary, then probabilistic classification methods might be used (eg. penalized regression, neural network, tree-based methods like random forests or boosted trees, etc...). In the following discussion we will often leverage predictive models $\widehat{\mu}_Y(x)$ and $\widehat{\mu}_Z(x)$, and care must be taken in estimating these models (using various statistical/machine learning tools, with proper selection of tuning parameters via split-sample validation, etc...). There is an enormous literature on building such models that we cannot hope to engage with here. However, we note that our ability to leverage these ideas in evaluating dependence is a strong asset for our method. Our asymptotic results will tend to rely on the following assumption:

Assumption 2.1. *Suppose we have n observations $o_i = (x_i, y_i, z_i)$, $i = 1, \ldots, n$ drawn iid from some distribution P. Let $\widehat{\mu}_Y$ and $\widehat{\mu}_Z$ be estimators of $\mu_{P,Y}$, $\mu_{P,Z}$ based on those observations. We assume that those estimators each fall in a P-Donsker Class [115], and further that*

$$\int [\widehat{\mu}_Y(x) - \mu_{P,Y}(x)]^2 dP(x) = o_p\left(n^{-1/2}\right),$$
$$\int [\widehat{\mu}_Z(x) - \mu_{P,Z}(x)]^2 dP(x) = o_p\left(n^{-1/2}\right).$$

This is just saying that our predictive models converge to the truth sufficiently fast. For correctly specified low/moderate dimensional parametric models (eg. linear/logistic regression) this will be satisfied (in fact the rate is actually $O_p\left(n^{-1}\right)$). This will also be the case for various nonparametric and high-dimensional methods under fairly general assumptions including the Lasso [104], additive models [88], and neural network models [5].

From here we can consider estimating our dependence measures. We begin with the

expected conditional covariance $\Psi_1(P)$. In this case we propose a natural plug-in estimator:

$$\widehat{\Psi}_1 \equiv \frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{\mu}_Y(x_i)][z_i - \widehat{\mu}_Z(x_i)], \tag{2.6}$$

in which we use our predictive models $\widehat{\mu}_Y$ and $\widehat{\mu}_Z$. As discussed in the next theorem this estimator is quite well-behaved.

Theorem 2.1. *Suppose Assumption 2.1 holds for $\widehat{\mu}_Y$ and $\widehat{\mu}_Z$. Then the plug-in estimator $\widehat{\Psi}_1$ is $\sqrt{n}$-consistent, asymptotically linear, and nonparametrically efficient with influence function $D_P^{(1)}(o_i) = (y_i - \mu_{P,Y}(x_i))(z_i - \mu_{P,Z}(x_i)) - \Psi_1(P)$. This additionally implies that $\widehat{\Psi}_1$ is asymptotically normal:*

$$\sqrt{n}[\widehat{\Psi}_1 - \Psi_1(P)] \to_d N[0, \sigma_1^2(P)], \tag{2.7}$$

where $\sigma_1^2(P) = \int [D_P^{(1)}(o)]^2 dP(o)$.

It is straightforward to obtain a consistent estimator of the asymptotic variance $\sigma_1^2(P)$, which is $\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left([y_i - \widehat{\mu}_Y(x_i)][z_i - \widehat{\mu}_Z(x_i)] - \widehat{\Psi}_1\right)^2$. This can be used with asymptotic-normality to form confidence intervals for Ψ_1 with asymptotically correct coverage. In addition, we should note that, so long as Assumption 2.1 holds, the plug-in estimator $\widehat{\Psi}_1$ has the same first-order behaviour (rate and variance), as the plug-in estimator with $\mu_{P,Y}$ and $\mu_{P,Z}$ known (which is first-order optimal in that case). This means that, under Assumption 2.1, there is no asymptotic cost to estimating the predictive models. These results can be shown by simple calculation (see supplementary materials).

We will postpone a discussion of estimating Ψ_2, and first discuss estimation of Ψ_3. We use a similar plug-in for Ψ_3: $\widehat{\Psi}_3 = \frac{\widehat{\Psi}_1}{\sqrt{\widehat{V}_Y \widehat{V}_Z}}$, where $\widehat{V}_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu}_Y(x_i))^2$ and $\widehat{V}_Z = \frac{1}{n}\sum_{i=1}^{n}(z_i - \widehat{\mu}_Z(x_i))^2$. Using a similar direct calculation, we can show that $\widehat{V}_Y$ and $\widehat{V}_Z$ are asymptotically linear and efficient estimates of $V_Y(P)$ and $V_Z(P)$. Thus, by applying the delta-method we get the following result

Theorem 2.2. *Suppose Assumption 2.1 holds for $\widehat{\mu}_Y$ and $\widehat{\mu}_Z$. Then the plug-in estimator $\widehat{\Psi}_3$ is $\sqrt{n}$-consistent, asymptotically linear, and nonparametrically efficient with influence function $D_P^{(3)}(o_i) = \frac{(y_i - \mu_{P,Y}(x_i))(z_i - \mu_{P,Z}(x_i))}{\sqrt{V_Y(P)V_Z(P)}} - \Psi_3(P)\left[\frac{(y_i - \mu_{P,Y}(x_i))^2}{2V_Y(P)} + \frac{(z_i - \mu_{P,Z}(x_i))^2}{2V_Z(P)}\right]$. This additionally implies that $\widehat{\Psi}_3$ is asymptotically normal:*

$$\sqrt{n}[\widehat{\Psi}_3 - \Psi_3(P)] \rightarrow_d N[0, \sigma_3^2(P)], \tag{2.8}$$

where $\sigma_3^2(P) = \int [D_P^{(3)}(o)]^2 dP(o)$.

We can similarly use a consistent estimate of $\sigma_3^2(P)$, and combine that with asymptotic normality to build a confidence interval for Ψ_3. This again has the same efficiency as the optimal estimator with $\mu_{P,Y}$ and $\mu_{P,Z}$ known.

Building an estimator for $\Psi_2(P)$ is a bit more complicated. Here, we must analyze the canonical gradient of $\Psi_2(P)$ under a nonparametric model. This informs us about the low-order terms in a von-mises expansion, and allows us to calculate the so-called "one-step" correction needed to update our plug-in estimator to construct an efficient estimator [14]. In order to follow this path, we also need estimators of $\mathrm{Cov}(Y, Z|X = x)$, $\sigma_Y^2(x)$, and $\sigma_Z^2(x)$. We will denote such estimators by $\widehat{\mathrm{Cov}}(Y, Z|X = x)$, $\widehat{\sigma}_Y^2(x)$, and $\widehat{\sigma}_Z^2(x)$. Coming up with strong estimators for these intermediate quantities is a significant hurdle in estimating Ψ_2 well, and a major reason why we instead propose Ψ_3 as a standardized measure of conditional dependence.

Based on all of this, the estimator we propose for $\Psi_2(P)$ is $\widehat{\Psi}_2 = \widetilde{\Psi}_2 + \frac{1}{n}\sum_{i=1}^n \widetilde{D}^{(2)}(o_i)$, where $\widetilde{\Psi}_2$ is a naive estimator of form

$$\widetilde{\Psi}_2 = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{(y_i - \widehat{\mu}_Y(x_i))(z_i - \widehat{\mu}_Z(x_i))}{\sqrt{\widehat{\sigma}_Y^2(x_i)\widehat{\sigma}_Z^2(x_i)}} \right\} \tag{2.9}$$

and

$$\widetilde{D}^{(2)}(o_i) = \left[\frac{\widehat{\mathrm{Cov}}(Y, Z|X = x_i)}{\sqrt{\widehat{\sigma}^2_Y(x_i)\widehat{\sigma}^2_Z(x_i)}}\right] \times \left[\frac{(y_i - \widehat{\mu}_Y(x_i))^2}{2\widehat{\sigma}^2_Y(x_i)} + \frac{(z_i - \widehat{\mu}_Z(x_i))^2}{2\widehat{\sigma}^2_Z(x_i)} - 1\right] \tag{2.10}$$

Here $D^{(2)}$ is the canonical gradient (or equivalently the efficient influence function) of $\Psi_2(P)$ in the nonparametric model-class.

Standard theory for such one-step estimators gives us the following result:

Theorem 2.3. *Suppose $\widehat{\mu}_Y(x)$, $\widehat{\mu}_Z(x)$ satisfy Assumption 2.1, and similarly estimators $\mathrm{Cov}(Y, Z|X = x)$, $\sigma^2_Y(x)$, and $\sigma^2_Z(x)$ are also from P-Donsker classes, and converge to the truth at that same $n^{-1/2}$ rate in squared error loss. Then the estimator $\widehat{\Psi}_2$ is $\sqrt{n}$-consistent, asymptotically linear, and nonparametrically efficient with influence function $D^{(2)}_P(o)$ defined in (2.10), This additionally implies that $\widehat{\Psi}_2$ is asymptotically normal:*

$$\sqrt{n}[\widehat{\Psi}_2 - \Psi_2(P)] \to_d N[0, \sigma^2_2(P)], \tag{2.11}$$

where $\sigma^2_2(P) = \int [D^{(2)}_P(o)]^2 dP(o)$.

Theorem 2.3 has requirements on convergence of additional intermediate quantities (conditional covariances and conditional variances). In practice, even in simple scenarios $\widehat{\Psi}_2$ performs much more poorly than $\widehat{\Psi}_1$ and $\widehat{\Psi}_3$. The theoretical route we took to derive this "efficient" estimator, could also have been applied for Ψ_1 and Ψ_2 to construct efficient estimators. It turns out, that in those cases, we would have ended up with *precisely* the plugins $\widehat{\Psi}_1$ and $\widehat{\Psi}_2$ from such constructions (however, one can more easily show efficiency of those estimators from direct calculation).

2.4.1 Double Robustness of $\widehat{\Psi}_1$

In Assumption 2.1, we give separate convergence rates bounds for each predictive model. In fact, for the result of Theorem 2.1 we only require that $R_1(\widehat{P}_n, P) \equiv \int [\widehat{\mu}_Y(x) - \mu_{P,Y}(x)][\widehat{\mu}_Z(x) -$

$\mu_{P,Z}(x)]dP(x) = O_P(n^{-1/2})$. In particular, this is precisely the second-order term from an asymptotic expansion of our estimator. Using this, we can directly show that our estimator $\widehat{\Psi}_1$ is *doubly robust* in that

- $\widehat{\Psi}_1$ is consistent if either one of $\widehat{\mu}_Y(x)$ and $\widehat{\mu}_Z(x)$ is consistent, and in a P-Glivenko-Cantelli Class. (and thus $R_1(\widehat{P}_n, P) = o_P(1)$)
- $\widehat{\Psi}_1$ is efficient if $\widehat{\mu}_Y(x)$ and $\widehat{\mu}_Z(x)$ converge sufficiently fast that $R_1(\widehat{P}_n, P) = o_P(n^{-1/2})$.

This indicates additional robustness of $\widehat{\Psi}_1$ to model misspecification [91, 113]. Even if one of $\widehat{\mu}_Z$ and $\widehat{\mu}_Y$ is inconsistent, $\widehat{\Psi}_1$ will still remain consistent as long as the other one is consistent. Unfortunately, neither the expected conditional correlation, nor the scaled expected conditional covariance estimators are double-robust. In particular, the scaled expected conditional covariance has second-order remainder terms associated with estimating each expected conditional variance which separately involve convergence of $\widehat{\mu}_Y(x)$ and $\widehat{\mu}_Z(x)$. See supplementary materials for details about remainder terms.

2.4.2 Suboptimal Estimators

To some degree, it is a happy coincidence that the estimators for Ψ_1 and Ψ_2 proposed in Section 2.4 are simple and turn out to be first-order optimal. Generally simple plug-in estimators will not even be rate optimal (and converge at a slower rate than $n^{-1/2}$). For example, one might consider an alternative representation of $\Psi_1(P) = \mathrm{E}_P[\mathrm{E}_P(YZ|X) - \mathrm{E}_P(Y|X)\,\mathrm{E}_P(Z|X)]$, and thus consider estimating $\Psi_1(P)$ by

$$\widehat{\Psi}_{1,naive} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_{YZ}(x_i) - \widehat{\mu}_Y(x_i)\widehat{\mu}_Z(x_i)\right], \tag{2.12}$$

where $\widehat{\mu}_{YZ}(x_i)$ is an estimator of $\mathrm{E}_P(YZ|X)$. If $\widehat{\mu}_Y$ and $\widehat{\mu}_Z$ do not converge at a parametric rate (of n^{-1} in MSE)– when using ML-based estimates they generally will not– $\widehat{\Psi}_{1,naive}$ will converge at slower than an $n^{-1/2}$ rate. One could similarly define a simple estimator of $\Psi_2(P)$,

$\widehat{\Psi}_{2,naive} = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{\text{Cov}}(Y,Z|x_i)}{\sqrt{\widehat{\sigma}^2_Y(x_i)\widehat{\sigma}^2_Z(x_i)}}$. Unfortunately, as in the case of $\widehat{\Psi}_{1,naive}$, this estimator will not be efficient or even converge at a $n^{-1/2}$ rate.

2.4.3 Constructing Confidence Intervals

Constructing a confidence interval based on the so-called naive estimators, $\widehat{\Psi}_{1,naive}$ and $\widehat{\Psi}_{2,naive}$, is difficult. Due to the excess bias, they are, in general, not asymptotically linear, so confidence intervals based on Gaussian approximations are not possible. In addition, resampling methods including bootstrapping, are generally invalid in this context. Fortunately, $\widehat{\Psi}_1$, $\widehat{\Psi}_2$ and $\widehat{\Psi}_3$ do not suffer from these issues. As shown in Theorems 2.1-2.3, these centered estimators converge in distribution to mean-zero normal variables with asymptotic variance $\sigma_j^2(P) = \int [D_P^{(j)}(o)]^2 dP(o)$ for $j = 1, 2, 3$. Thus, if we estimate our variances by

$$\widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{D}^{(j)}(o_i)^2, \tag{2.13}$$

where $\widehat{D}^{(j)}$ is any consistent estimator of the influence function, we can form valid confidence intervals. Then, by leveraging asymptotic normality, we can construct a $(1-\alpha)\%$ Wald-type confidence interval for Ψ_j as

$$[\widehat{\Psi}_j - n^{-1/2} q_{1-\alpha/2}\widehat{\sigma}_j, \widehat{\Psi}_j + n^{-1/2} q_{1-\alpha/2}\widehat{\sigma}_j], \tag{2.14}$$

which has asymptotically correct coverage. q_α stands for the α-th quantile of a standard normal distribution.

Asymptotic linearity can be leveraged more broadly to give intervals for multiple pairs of features with correct simultaneous coverage. In particular, suppose we are in an asymptotic regime with p fixed and n growing. Consider 2 pairs of features (j_1, j_2), and (j_3, j_4) with $j_1 \neq j_2$ and $j_3 \neq j_4$, (this can be extended to any number of pairs). In this case, we consider estimation of $\left[\Psi_1^{j_1,j_2}, \Psi_1^{j_3,j_4}\right]^\top$, the expected conditional covariance of both pairs of features.

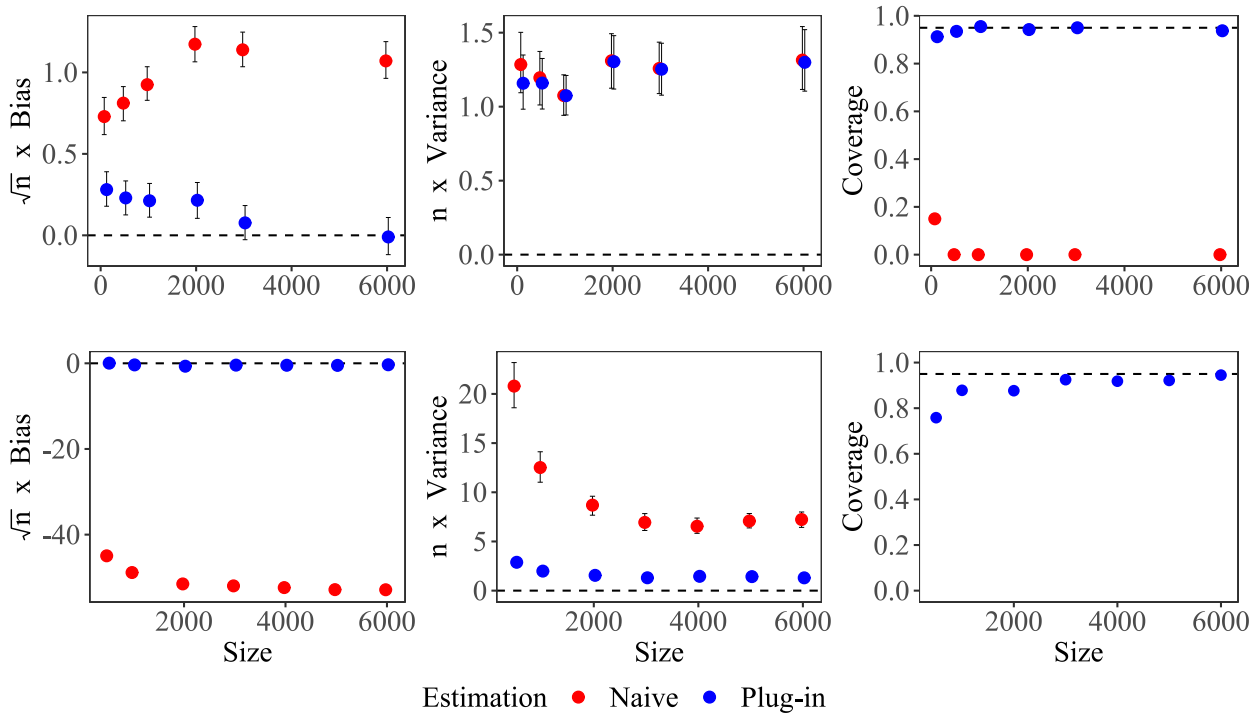

Figure 2.1: Empirical $\sqrt{n}-$scaled bias, Empirical $n-$scaled variance and empirical coverage of 95% confidence interval of $\widehat{\Psi}_1$ (blue) and $\widehat{\Psi}_{1,naive}$ (red) for the low-dimensional case (top) and the high-dimensional case(bottom). We only provide a bootstrap-based confidence interval for naive estimators in the low-dimensional case to show its failure.

Here, it is straightforward to show that under Assumption 2.1, we have

$$\sqrt{n}\left[\begin{pmatrix}\widehat{\Psi}_1^{j_1,j_2}\\ \widehat{\Psi}_1^{j_2,j_3}\end{pmatrix}-\begin{pmatrix}\Psi_1^{j_1,j_2}\\ \Psi_1^{j_2,j_3}\end{pmatrix}\right]\to N\left(0,\Sigma\right),$$

where Σ is defined based on expectations of products of influence functions for each estimator. This idea generalizes to arbitrary (but fixed) numbers of covariates, and can also be applied to estimation of Ψ_2, and Ψ_3. This joint normality can be combined with standard methods in multiple testing to construct confidence intervals with simultaneous coverage [112].

2.4.4 Relationship to De-biased Lasso

In addition to graphical modeling, other meaningful measures can be obtained by slightly modifying Ψ_1. One measure of particular interest is

$$\Phi = \frac{\mathrm{E}[\mathrm{Cov}(Y, Z|X)]}{\mathrm{E}[\mathrm{Var}(Z|X)]}. \tag{2.15}$$

Φ is a nonparametric functional, that combines expected conditional variance and covariance (similar to Ψ_1). In fact, as with Ψ_1, we can use a simple plug-in estimator (with estimated conditional means constructed using any suitable machine learning technique) to estimate and make inference for Φ. If we further assume that we are working in a parametric space and the data (Y, Z, X) are generated from a linear model $\mathrm{E}[Y|Z, X] = \gamma Z + \beta X$, Φ is precisely the coefficient γ [79]. In low dimensional problems γ is estimated efficiently by standard linear regression — in high-dimensional problems it is common to use the Lasso [104, 81] with de-biasing to conduct inference [90, 31]. The work in this chapter gives an alternative approach to estimation and inference. In particular, in the challenging case that the features are high-dimensional, the (theoretically optimal) plug-in estimator $\widehat{\Phi}$ is consistent and efficient (if the conditional mean estimates are sufficiently good). Under suitable conditions, the de-biased lasso will give an estimator with the same first order behavior when the design matrix is random [45]. However, the de-biased lasso requires estimation of Σ^{-1} (usually by node-wise regression) which our nonparametric approach does not. Thus, the results in this paper provide an alternative for obtaining the estimators and confidence intervals of regression coefficients for linear models with either low- or high-dimensional features.

2.5 Experiments

In this section, we assess the performance of the proposed (theoretically optimal) plug-in estimators of global dependence measures, in terms of the asymptotic performance, as well as their effectiveness in conditional independence testing and graph recovery. Here, we present the main results and provide additional results in supplementary materials.

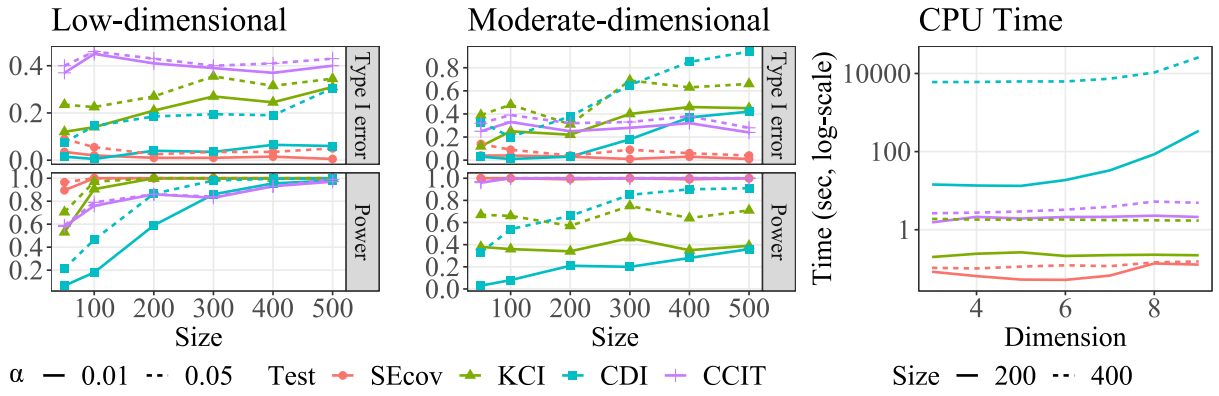

Figure 2.2: Left and middle: Type I error and power of three conditional independent testing methods for a low- and a moderate-dimensional case. Right: average CPU time taken by four tests. SEcov, KCI-test and CDI-test are all implemented in R. CCIT-test is implemented in Python.

2.5.1 Asymptotic Performance

We present the asymptotic properties of Ψ_1 by computing the empirical bias, variance, and coverage of 95% Wald-type confidence interval in the setting of low-, moderate, and high-dimensional features.

We start with a simple scenario, where the conditioning variable X is univariate:

$$Y = \sin(3X) + e_y, \quad Z = \cos(2X) + e_z, \tag{2.16}$$

where $X \sim$ Uniform(0,2) independent of $\vec{e} = (e_y, e_z)^T \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right]$.

Then, we consider a setting of high-dimensional features, where we generate Y and Z from a linear model:

$$Y = X\beta_y + e_y, \quad Z = X\beta_z + e_z, \tag{2.17}$$

where $X \sim N(0, I_{5000})$, $\beta_y = (\underbrace{1, ..., 1}_{10}, \underbrace{0, ..., 0}_{4990})$ and $\beta_z = (\underbrace{-1, ..., -1}_{10}, \underbrace{0, ..., 0}_{4990})$. The error term $\vec{e} = (e_y, e_z)^T$ is the same as in the low-dimensional case.

In both cases, the true values of Ψ_1 are -0.5. We generate random datasets of size

$n \in \{500, 1000, 2000, ..., 6000\}$ and estimate Ψ_1 (we run 400 simulates for each sample size). The conditional means are estimated by local polynomial regression in the low-dimensional case and by lasso algorithm in the high-dimensional case. We compare our (theoretically optimal) plug-in estimator $\widehat{\Psi}_1$ to the naive estimators: $\widehat{\Psi}_{1,naive}$ in (2.12).

Figure 2.1 shows that, the empirical $\sqrt{n}$-scaled bias of our theoretically optimal plug-in estimator $\widehat{\Psi}_1$ goes toward zero with increasing sample size, which corresponds to our asymptotic result. This is not the case for the naive estimator. The confidence interval of $\widehat{\Psi}_1$ converges to the nominal 95% as sample size increases. As expected, due to excess bias, the bootstrap interval based on the "naive" estimators performs poorly (with coverage actually converging to 0). See supplementary materials for experiments of a moderate-dimensional case and the evaluation of Ψ_3.

2.5.2 Conditional Independence Testing.

We examine the probabilities of Type I error under $Y \perp\!\!\!\perp Z|X$ and the power under $Y \not\perp\!\!\!\perp Z|X$. Here, we consider the scenarios where $X \in \mathbb{R}^1$ and $X \in \mathbb{R}^5$ respectively. We compare the test based on the scaled expected conditional covariance (SEcov), i.e. Ψ_3, with KCI-test [132], CDI-test [117] and CCIT-test [96]. The conditional means for Ψ_3 are estimated by local polynomial regression when $X \in \mathbb{R}^1$ and by random forest when $X \in \mathbb{R}^5$.

In the low-dimensional setting, we still use model (2.16) to generate the data (Y, Z, X). For type I error, we let $\mathrm{Cov}(e_y, e_z)^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ such that $Y \perp\!\!\!\perp Z|X$. For power, we let $\mathrm{Cov}(e_y, e_z)^T = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$, such that $Y \not\perp\!\!\!\perp Z|X$. In the moderate-dimensional setting, we use the same pattern as the Case1 in [132] for comparison. Y and Z are generated by $G(F(X) + E)$, $X \in \mathbb{R}^5$, where G and F are mixtures of linear, cubic, and tanh functions and are different for Y and Z. E is independent with both Y and Z. Under this mechanism, $Y \perp\!\!\!\perp Z|X$ holds. For $Y \not\perp\!\!\!\perp Z|X$, we add errors $\cosh(e_y)$ to Y and $\cosh(e_z^2)$ to Z where $e_y, e_z \sim_{\text{iid}} N(0, 1)$.

Figure 2.2 shows that Ψ_3 is always capable of controlling type I errors and achieving a high power, regardless of the dimension of the conditioning set. However, this is not the case for other tests. When $X \in \mathbb{R}^1$, the power of KCI- and CDI-test gradually increases with increasing sample size. They can control type I errors at a relatively low level but not comparable to the performance of Ψ_3. When $X \in \mathbb{R}^5$, both kernel-based tests collapse. That is, they almost always reject the null hypothesis when $Y \perp\!\!\!\perp Z|X$, and often fail to reject the null when $Y \not\perp\!\!\!\perp Z|X$. The CCIT-test achieves a relatively high power but struggles to control type type-I errors in both low- and moderate-dimensional settings. In addition, the CDI test is much less efficient compared to the other three. With regard to computation, estimating Ψ_3 is the most efficient method for each fixed sample size, since it only requires the estimation of mean models.

2.5.3 Graph Recovery.

We now attempt to reconstruct the graph using SEcov, i.e. Ψ_3, with moderate dimensional features (the conditional means are estimated by random forest). We make comparison with Gaussian graphical model (GGM), and transelliptical graphical model (TGM) [71] where the CLIME estimator [19] using Kendall's taus is employed. The graphs are generated from the following cases:

- Case1 (Gaussian): $X \sim N_8(0, \Sigma)$.

- Case2 (Copulas): $Z \sim N_8(0, \Sigma)$, $U = \Phi(Z)$, $X_i = f_i^{-1}(U_i)$ where f_i^{-1} are quantile functions of Gamma$(2,1)$, Gamma$(2,1)$, Beta$(2,2)$, Beta$(2,2)$, t(5), t(5), Unif$(0,1)$, and Unif$(0,1)$ for $i = 1, ..., 8$.

- Case3 (Transelliptical): $X \sim TE_8(\Sigma, \xi; f)$. $\xi \sim \chi_p$ and $f = \{f_1, ..., f_8\} = \{h_1, h_2, h_3, h_4, h_1, h_2, h_3, h_4\}$, where $h_1^{-1}(x) = \sqrt{\exp(x)}$, $h_2^{-1}(x) = \text{sign}(x)|x|^{1/2}$, $h_3^{-1}(x) = x^3$, and $h_4^{-1}(x) = \Phi(x)$.

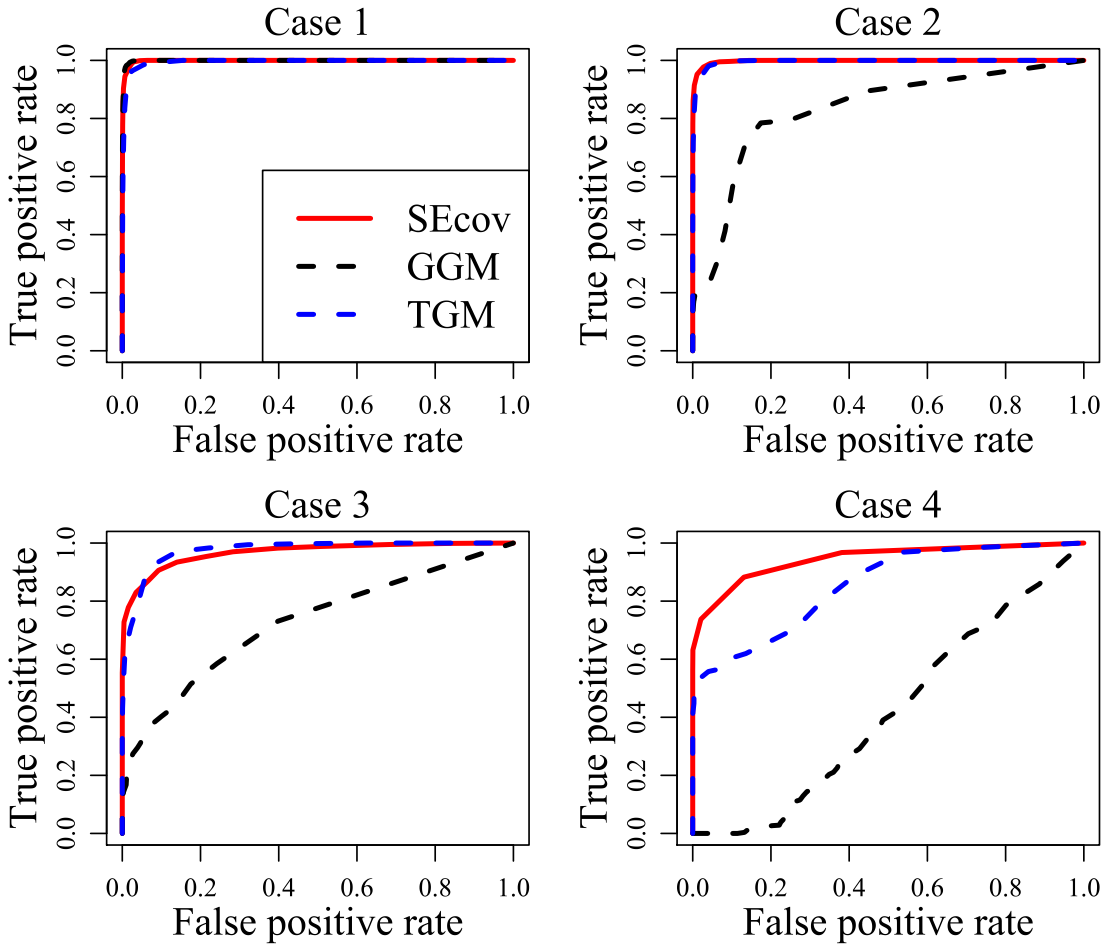

Figure 2.3: ROC curves of graph recovery for different methods in Case1-Case4. $n = 400$, $p = 8$.

- Case4 (non-Gaussian, non-copulas, non-transelliptical): $X_1 = X_2 + X_3 + X_4/2 + \sin(X_5) + X_6^2 + \exp(X_7) + X_8$ and $X_2 = \sin(X_7) + |X_8|$, where $X_3, ..., X_8 \sim_{\text{iid}} \exp(2)$.

Figure 2.3 shows that, all three methods work extremely well only when the data is Gaussian distributed (Case1). When the data follows a copulas (Case2) or transelliptical distribution (Case3), both TGM method and Ψ_3 have a comparably great performance while GGM become much less effective due to the model misspecification. We note that, if the data has a highly skewed transelliptical distribution, Ψ_3 may work poorly and TGM remains valid. For Case 4 where the data is non-Gaussian, non-copulas, and non-transelliptical, GGM

method totally collapses, which is almost equivalent to a coin flip. The effectiveness of TGM method is also compromised since it uses a misspecified model. On the contrary, Ψ_3 which does not depend on any model assumptions still presents a strong performance.

2.6 Discussion

In this paper, we introduce three global measures for evaluating conditional dependence and reconstructing a conditional dependence graph. These measures are model-agnostic and we show that there exist natural and simple plug-in estimators that are asymptotically normal and efficient under mild conditions. Thus, we can construct Wald-type confidence intervals with asymptotically correct coverage. These tasks have proven difficult for existing graphical modeling methods.

One major strength of this work is in that the estimation of the proposed global measures only requires estimating two conditional mean models. Our framework allows us to use flexible machine learning tools for these estimates. Thus, the efficacy of our methodology is intimately connected to our ability to build a good predictive model: If we can build effective predictive models, our methodology can leverage that, and should do a good job evaluating conditional independence. This means, as the field's ability to engage in predictive modeling grows, so will the scope of this methodology. For example, in the high-dimensional setting, one might use Lasso, or tree-based ensembles to regress out the conditioning variables. If the conditioning variables take form of images, or text documents, one could use deep-learning (with enough data) for that adjustment. The predictive methodology can and should be selected to fit the context.

People may concerned about the effectiveness of the proposed methodology in very high-dimensional settings, as it requires fitting $\sim p^2$ models. However, since each conditional mean is estimated independently, the dependence between every pair of features can be evaluated entirely in parallel. Additionally, one might also consider adopting some form of "pre-screening". For example, one may apply a simpler method (with potential false positives) first to create a network with a super-set of edges and then deploy the methodology proposed

in this chapter to refine this to a more accurate graph.

Chapter 3

NONPARAMETRIC INFERENCE UNDER THE PRESENCE OF LINEARITY

3.1 Introduction

In science, it is often of interest to understand the relationship between a single outcome and a number of explanatory variables. For example, researchers try to identify common genetic variants that are associated with disease phenotype [74, 17, 92], to establish association between voxels of brain images and clinical variables such as age and gender [16, 133], and to understand how the housing prices change with factors in real estate [70, 83].

Regression analysis is the most prominent family of techniques for addressing these problems, where parametric linear regression plays a primary role and provides arguably the simplest and the most interpretable tool to estimate the relationship between the target outcome and other features. Suppose that we have n pairs of observations $\left\{(y^{(i)}, \mathbf{x}^{(i)})\right\}_{i=1}^{n}$ that are independently drawn from a linear model:

$$Y = \mathbf{X}^T\boldsymbol{\beta} + \epsilon, \tag{3.1}$$

where $Y \in \mathbb{R}$ represents the outcome and $\mathbf{X} = (X_1, ..., X_p)^T \in \mathbb{R}^p$ are random covariates representing features. We assume the error term ϵ is independent of $\mathbf{X}$, with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$ denote the vector of unknown regression coefficients. From (3.1), one can see that β_j encodes the "relationship between the outcome Y and the covariate X_j": Y is linearly associated with X_j, and a one unit increase in X_j (with all other X fixed) would lead to a β_j increase in Y.

To estimate and make inference on these unknown $\boldsymbol{\beta}$, it is standard to ordinary least

squares (OLS) regression, by minimizing the Euclidean distance between Y and the space spanned by the columns in $\mathbf{X}$. It is well known that that the OLS estimator has favorable theoretical guarantees if the linear model (3.1) truly holds: It is the best linear unbiased and $\sqrt{n}$-consistent estimator [76, 94, 120]. In particular, when the design matrix is fixed, [64] pointed out that OLS method may still work under model misspecification. However, in that case, OLS may lose the interpretability. When the design matrix is random, then OLS method requires the true parametric model to be correctly specified and the covariance matrix $\mathrm{E}(\mathbf{X}^T\mathbf{X})$ to be non-singular [39, 1, 62].

To deal with the case when $p > n$, penalized techniques have been proposed, e.g., lasso regression [104], ridge regression [52], and the elastic net [135]. However, none of these produces an asymptotically unbiased estimator of $\boldsymbol{\beta}$. This makes inference quite challenging: For example the bootstrap fails due to lack of continuity of the estimator's limiting distribution. To remedy this, [131] and [111] proposed a de-biased lasso algorithm, which yields a non-sparse estimator. Under certain regularity conditions and sparsity assumptions, it has been shown that the estimator is asymptotically linear and normal. Using this, one can construct asymptotically valid confidence intervals and performance valid statistical tests.

These existing methods, however, are all based on the strong parametric specification of the data generating mechanism: They assume that Y is truly linearly associated with the components of $\mathbf{X}$. In many real-world applications, people are primarily interested in a few of the explanatory features, e.g., interventions or exposures, while controlling for other potential confounders (which correspond to nuisance parameters here). In this case, using the model given in (3.1) which imposes a linear relationship between the outcome and potential confounders, might be unnecessarily restrictive. Therefore, in this work, we try to relax that assumption and consider the following parameter for measuring the "adjusted linear association" between any two variables Y and X_j, under a fully unrestricted *non-parametric* model space,

$$\Phi = \frac{\mathrm{E}[\mathrm{Cov}(Y, X_j|\mathbf{X}_{-j})]}{\mathrm{E}[\mathrm{Var}(X_j|\mathbf{X}_{-j})]}, \tag{3.2}$$

for random vector $\mathbf{O} = (Y, \mathbf{X})$ with $Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, ..., X_p) \in \mathbb{R}^p$ from an unknown distribution P. $\mathbf{X}_{-j} = \mathbf{X} \setminus X_j$. In (3.2), $\text{Cov}(Y, X_j|\mathbf{X}_{-j})$ is the conditional covariance of Y and X_j given $\mathbf{X}_{-j}$, and $\text{Var}(X_j|\mathbf{X}_{-j})$ is the conditional covariance of X_j given $\mathbf{X}_{-j}$.

Above parameter is reminiscent of the OLS estimator for simple linear regression i.e., $p = 1$ in model (3.1), which is calculated by the ratio of the sample covariance between Y and X_j and the sample variance of X_j. In this sense, Φ might be a reasonable parameter to quantify linear association. In addition, Φ has several favorable properties: (i) it allows one to adjust for any linear or nonlinear confounding effects of any other variables, and thus measures the conditionally linear association; (ii) it does not implicitly assume any parametric form for the data generating mechanism: It is just a functional (averaging all the local quantities) that maps the joint distribution P to a real number; (iii) it does not have constraints on the dimension of the conditioning set and might accommodate low- or high-dimensional data.

Estimation of Φ is simple and straightforward. We will show that estimation of Φ primarily depends on the estimation of the following conditional means,

$$\mu_{P,Y}(\boldsymbol{x}_{-j}) := \text{E}(Y|\mathbf{X}_{-j} = \boldsymbol{x}_{-j}) \quad \text{and} \quad \mu_{P,j}(\boldsymbol{x}_{-j}) := \text{E}(X_j|\mathbf{X}_{-j} = \boldsymbol{x}_{-j}). \tag{3.3}$$

Thus, the problem of evaluating the linear association reduces to a *canonical prediction problem*, which enables us to naturally leverage flexible machine learning techniques.

Although we propose Φ (and its corresponding estimator) for use in a nonparametric model space, it is of interest to compare it to the parameter (and estimator) indicated by OLS (and its de-biased lasso extension in high-dimensional linear regression problems). Specifically, we shall consider the parameter Φ in the context of a correctly specified linear model and address the following questions: (i) how can we interpret the parameter there; (ii) How does the natural estimator of Φ compare to the OLS estimator in this problem (iii) What intuition do we have for Φ when the linear regression model is misspecified.

In the rest of this chapter, we shall first formally introduce the target parameter in the form of functional in a nonparametric space and connect it to the regression coefficients under

the presence of linearity. Then, in Section 3.2, we will provide a simple and natural estimator that enjoys favorable asymptotic properties in general. Following that, in Section 3.3, we assume the linear model is correctly specified and compare the proposed estimator with OLS estimation and de-biased lasso estimation in low- and high-dimensional settings (with regard to consistency, efficiency and robustness). Then, we conduct experimental studies to empirically explore finite sample behavior and verify these theoretical results in Section 3.4. Finally, we end this chapter with a discussion.

3.2 A Nonparametric Parameter for Conditionally Linear Association

Let $\boldsymbol{O} = (Y, \mathbf{X})$ denote a random vector drawn from an unknown distribution $P \in \mathcal{P}$, where $Y \in \mathbb{R}$ and $\mathbf{X} = (X_{\cdot 1}, \ldots, X_{\cdot p})^T \in \mathbb{R}^p$. Here, $\mathcal{P}$ is an unrestricted, or nonparametric model space. The primary goal is to quantify the linear association between any one of those features, X_j, and the outcome Y. We define our target parameter as follows:

$$\begin{aligned} \Phi_j(P) &= \frac{\mathrm{E}_P[\mathrm{Cov}_P(Y, X_j|\mathbf{X}_{-j})]}{\mathrm{E}_P[\mathrm{Var}_P(X_j|\mathbf{X}_{-j})]} \\ &= \frac{\int [y - \mu_{P,Y}(\boldsymbol{x}_{-j})][x_j - \mu_{P,j}(\boldsymbol{x}_{-j})] dP(\boldsymbol{o})}{\int [x_j - \mu_{P,j}(\boldsymbol{x}_{-j})]^2 dP(\boldsymbol{o})}, \end{aligned} \tag{3.4}$$

for all $j = 1, ..., p$ where $\mathbf{X}_{-j}$ represents all covariates except for X_j, $\mu_{P,Y}(\boldsymbol{x}_{-j})$ and $\mu_{P,j}(\boldsymbol{x}_{-j})$ are as defined in (3.3). As mentioned in Section 3.1, the parameter $\Phi_j(P)$ is perhaps a reasonably intuitive measure to assess the conditional linear association since it is analogous to the simple OLS estimator while adjusting for the confounding effects of other covariates $\mathbf{X}_{-j}$: The numerator is the expected conditional covariance which is a measure of the average conditional dependence between Y and X_j; The denominator is the expected conditional variance, which serves as a scaling factor. Hence, the sign of Φ_j would tell whether, on average, there is a positive or negative association between Y and X_j: A positive Φ_j indicates that as the value of the explanatory variable increases, the outcome also tends to increase after controlling for the effects of other covariates. The magnitude of Φ_j signifies the strength

of such a trend: A greater magnitude of Φ_j implies that Y and X_j are more dependent.

The parameter $\Phi_j := \Phi_j(P)$ takes into account potential confounding effects by the remaining features $\mathbf{X}_{-j}$. Unlike linear regression modeling, there is no assumption that these confounding features are linearly associated with the outcome; and more generally there is no parametric specification of the conditional means $\mu_{P,Y}$ and $\mu_{P,j}$. This permits more flexibility of our method over classical linear regression.

We note that X_j here does not necessarily have to be continuous. In fact, when X_j is binary, it turns out that the proposed parameter Φ_j corresponds to a measure which can play a more causal role, as stated in the following proposition.

Proposition 3.2. *If X_j is a binary variable and has $X_j \sim$ Bernoulli(μ_j), where $\mu_j := \mathrm{E}(X_j|\mathbf{X}_{-j}) = p(X_j = 1|\mathbf{X}_{-j})$, the parameter Φ_j in (3.4) reduces to*

$$\Phi_j(P) = \mathrm{E}\left\{ \frac{\mu_j(1-\mu_j)}{\mathrm{E}[\mu_j(1-\mu_j)]} \left(\mathrm{E}[Y|X_j = 1, \mathbf{X}_{-j}] - \mathrm{E}[Y|X_j = 0, \mathbf{X}_{-j}]\right) \right\}, \tag{3.5}$$

which is the variance weighted average effect of $X_j = 1$, compared to $X_j = 0$.

The derivation is given in Appendix B.1. The result in Proposition 3.2 has also been noted in [86]. There, they briefly consider our parameter Φ_j, but only in the context of binary exposure (binary X_j), and relate it to the causal literature.

3.2.1 $\Phi_j(P)$ under the Presence of Linearity

Unlike coefficients in parametric models, Φ_j is defined nonparametrically, as a functional that only depends on the data generating mechanism. However, the construction of $\Phi_j(P)$ implies that it intuitively connects with the regression coefficients in a linear model. To see this, here we rewrite model (3.1) by

$$Y = X_j\beta_j + \mathbf{X}_{-j}^T\boldsymbol{\beta}_{-j} + \epsilon, \quad \epsilon \sim_{i.i.d.} (0, \sigma^2), \tag{3.6}$$

and further suppose that $\mathbf{X} \sim (\mathbf{0}, \Sigma_p)$, where the inverse of Σ_p is assumed to exist, i.e., $\Theta_p := \Sigma_p^{-1}$. Under linearity, we say that data $\mathbf{O}$ has a joint distribution $P_{\text{lm}} \in \mathcal{P}_{\text{lm}} = (\mathcal{P}_X, \mathcal{P}_Y, \mathcal{P}_\epsilon)$. $\mathcal{P}_{\text{lm}}$ represents a model space where the linear relationship $\mathrm{E}_{P_{\text{lm}}}(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$ always holds. For the remainder of this section, we will assume that linearity holds (that our joint distribution lies in $\mathcal{P}_{\text{lm}}$).

To remove the impact of nuisance piece, i.e., $\mathbf{X}_{-j}^T\boldsymbol{\beta}_{-j}$ when the linear model holds, we take conditional expectations with respect to $\mathbf{X}_{-j}$ on both sides of (3.6), and subtract the resulting expression from (3.6). This gives us:

$$Y - \mathrm{E}(Y|\mathbf{X}_{-j}) = [X_j - \mathrm{E}(X_j|\mathbf{X}_{-j})]\beta_j + \epsilon. \tag{3.7}$$

This "nonparametric projection" strategy has been widely used in the semiparametric inference literature [33, 68, 134]. When our joint distribution, P_{lm} lies in $\mathcal{P}_{\text{lm}}$, it follows directly from (3.7) that

$$\beta_j = \Phi_j(P_{\text{lm}}) \qquad \text{for } j = 1, ..., p. \tag{3.8}$$

The equivalence in (3.8) further illustrates that the proposed nonparametric parameter $\Phi_j(P)$ is a reasonable measure of conditional linear association. In particular, under a linear model space $\mathcal{P}_{\text{lm}}$, $\Phi_j(P_{\text{lm}})$ has the same interpretation as the regression coefficient β_j: One unit increase in X_j would result in β_j average increase in the outcome Y if the values of all other X_{-j} are fixed.

To estimate $\boldsymbol{\beta}$, it is standard to use OLS estimation [94] when $p < n$ and lasso-based methods [104] when $p > n$. To give meaningful results, both of these methods require the data generating mechanism to lie in $\mathcal{P}_{\text{lm}}$: The outcome of interest should be linearly associated with every covariate we measure $\mathbf{X}$. This parametric specification is unnecessarily restrictive when only the impact of a few features is of interests. Here, the relationship in (3.8) gives us an alternative: As long as we can find an efficient estimator for $\Phi_j(P)$, we are able to build an estimator and make inference for a parameter that reduces to the regression coefficient β_j, but permits more flexibility and robustness. In the next section, we will give more details

about estimation of this parameter.

3.2.2 Estimation procedure

To simplify exposition, we begin by introducing some notation: Let $\Psi_j(P) := \mathrm{E}_P[\mathrm{Cov}_P(Y, X_j|\mathbf{X}_{-j})]$ and $V_j(P) := \mathrm{E}_P[\mathrm{Var}_P(X_j|\mathbf{X}_{-j})]$. $V_j(P)$ is very similar to $\Psi_j(P)$ just replacing Y by X_j. In Chapter 2 and [125], we gave a simple and natural plug-in estimator for Ψ_j:

$$\widehat{\Psi}_j = \frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} - \widehat{\mu}_Y(\mathbf{x}_{-j}^{(i)})\right]\left[x_j^{(i)} - \widehat{\mu}_j(\mathbf{x}_{-j}^{(i)})\right]. \tag{3.9}$$

given observations $\{\mathbf{o}^{(i)}\}_{i=1}^n = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^n$ where $\mathbf{x}^{(i)} = (x_1^{(i)}, ..., x_p^{(i)})^T$ represents the ith realization of the covariates $\mathbf{X}$ and $y^{(i)}$ is the ith realization of Y. Here $\widehat{\mu}_Y\left(\mathbf{x}_{-j}^{(i)}\right)$ is an estimator of the true conditional mean of Y given $\mathbf{X}_{-j} = \mathbf{x}_{-j}^{(i)}$, i.e., $\mu_{P,Y}\left(\mathbf{x}_{-j}^{(i)}\right)$.

The form of (3.9) implies that, the performance of $\widehat{\Psi}_j$ depends on how well those conditional means can be estimated. In Chapter 2 and [125], we prove that $\widehat{\Psi}_j$ is *asymptotically linear and efficient* when $\widehat{\mu}_Y$ and $\widehat{\mu}_j$ converge to their truths at a sufficiently fast rate. Thus, the problem of estimating $\Psi_j(P)$ can be reduced to the canonical "predictive modeling" problem. This allows us to leverage many flexible machine learning techniques, including generalized additive models [49], local polynomial regression [95], random forests [69] etc., and even potentially make inference even when X is high-dimensional [104, 75]. Full discussion of this can be see in Chapter 2.

Here, we take the advantage of the fact that $\Phi_j(P)$ is a ratio of $\Psi_j(P)$ and $V_j(P)$, and propose the following estimator for $\Phi_j(P)$:

$$\widehat{\Phi}_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} - \widehat{\mu}_Y(\mathbf{x}_{-j}^{(i)})\right]\left[x_j^{(i)} - \widehat{\mu}_j(\mathbf{x}_{-j}^{(i)})\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[x_j^{(i)} - \widehat{\mu}_j(\mathbf{x}_{-j}^{(i)})\right]^2}, \tag{3.10}$$

by plugging in $\widehat{\Psi}_j$ and $\widehat{V}_j$ for Ψ_j and V_j, respectively. Later, we shall show that this estimator also has favorable theoretic guarantees: It is *asymptotically linear and efficient* under mild

conditions.

We note that $\widehat{\Phi}_j$ in (3.10) is obtained without assuming any parametric models, and is simply an estimator of $\Phi_j(P)$ regardless of the relationship between Y and $\mathbf{X}$. However, when data $\{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^n$ are truly generated from a linear model (3.6), this estimator $\widehat{\Phi}_j$ can be an alternative to the OLS and de-biased lasso estimators of the regression coefficient β_j, due to the equivalence $\beta_j = \Phi_j(P_{\text{lm}})$. However, by allowing more general non-parametric and/or machine learning techniques to regress out potential confounders, $\Phi_j(P)$ and its estimator $\widehat{\Phi}_j$ can give more meaningful summaries of relationships of interest when linearity does not hold.

3.2.3 Misspecification of the Linear Model

In many cases, we are primarily interested in the conditional association of only a few of our measured features with the outcome. In those cases, we still need to adjust for other variables to remove any potential confounding effect. In such a setting, the linear model in (3.6) is an unnecessarily restrictive assumption. Even in a classical statistical setting (wherein we think of parameters as indexing rather than summarizing a model), one need not assume a linear relationship between confounders and outcome. The *partially linear* model is a generalization of a linear model that codifies this:

$$Y = X_j\beta_j + g(\mathbf{X}_{-j}) + \epsilon. \tag{3.11}$$

(3.11) is a semiparametric model where Y is only assumed to be linearly associated with X_j and $g(\cdot)$ is a nonparametric component which could be either linear or nonlinear. When the data generating mechanism (P_{pl}) is given by such a partially linear model (in (3.11)), then we still have the identification $\Phi_j(P_{pl}) = \beta_j$, precisely as we did in the linear model. This equivalence is straightforward to show using the same "nonparametric projection" strategy. Since (i) Our estimator $\widehat{\Phi}_j$ in (3.10) is completely agnostic to whether linearity (or partial linearity) holds; and (ii) its asymptotic guarantees are derived without any parametric requirements (we will show this later), it follows that $\widehat{\Phi}_j$ remains effective for estimating β_j

under model (3.11): Its theoretical properties, i.e., consistency and asymptotic normality, will be maintained.

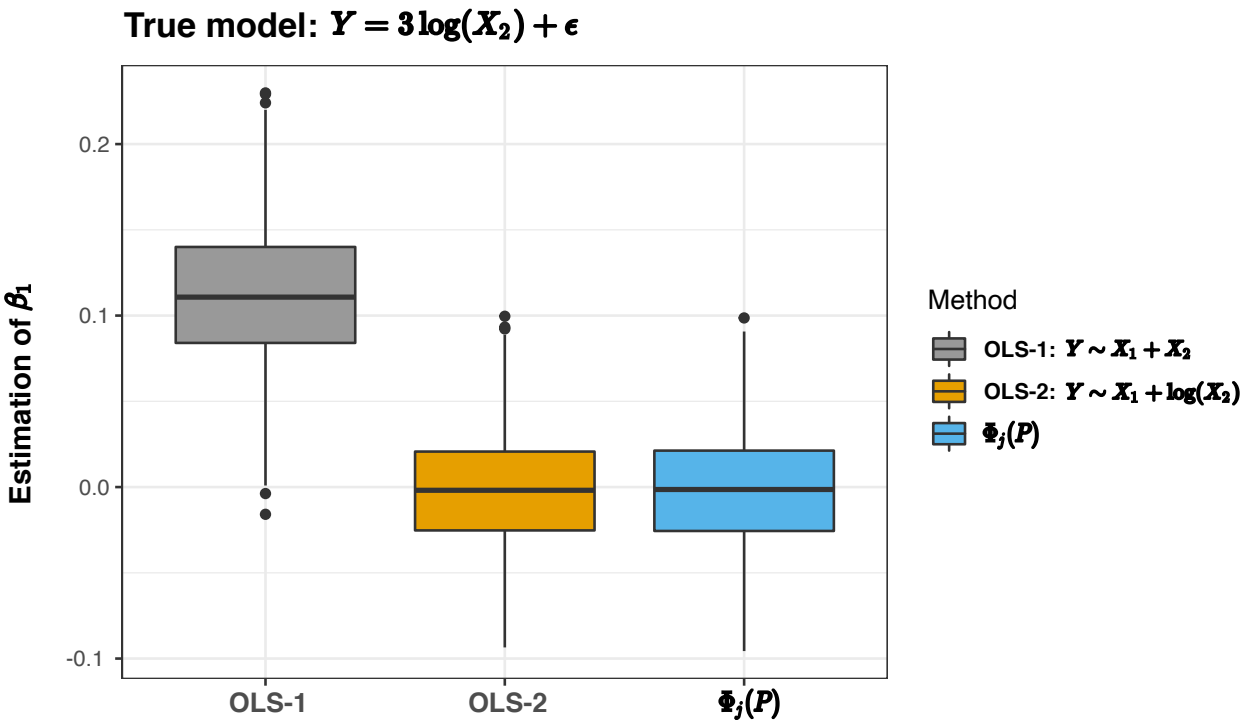

Figure 3.1: A simple illustration of the performance of OLS estimator and $\widehat{\Phi}_j$. Data are generated as $(X_1^*, X_2^*) \sim \mathbb{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = -0.6$. Then the observed covariates are $\mathbf{X} = (X_1, X_2) = (X_1^*, \exp(X_2^*))$. The true model is $Y = \beta_1 X_1 + \beta_2 \log(X_2) + \epsilon$ with $\beta_1 = 0$, $\beta_2 = 3$, and $\epsilon \sim N(0, 1)$. We apply OLS estimation and the proposed $\widehat{\Phi}_j(P)$ to estimate β_1 (in this case, $\beta_j = \Phi_j(P)$). For OLS, we consider two models: A incorrectly specified model OLS-1: $Y \sim X_1 + X_2$ and a correctly specified model OLS-2: $Y \sim X_1 + \log(X_2)$. For $\widehat{\Phi}_j$, the conditional means are estimated by local polynomial regression [55]. We generate 500 random data sets of size $n = 1000$.

However, this is not the case for other model-based approaches for estimating β_j, like OLS and de-biased lasso estimation. When the linear model is misspecified, OLS estimation could completely collapse. To illustrate this, we consider a simple case where Y is nonlinearly associated with X_2 and has no association with X_1 at all. Figure 3.1 shows that, when there

exists model misspecification, OLS-1 detects a spurious association between Y and X_1 as the estimated β_1 deviates from 0. In contrast, when the model is correctly specified as in OLS-2, the OLS estimator of β_1 is close to 0. The non-parametric form of $\Phi_j(P)$ allows us to accommodate this non-linear confounding, and apply the estimator $\widehat{\Phi}_j$ in (3.10). We see that $\widehat{\Phi}_j$ has comparable performance to OLS-2 in terms of the empirical bias and variance of the estimator. In practice, the performance of OLS-2 can cannot easily achieved by using OLS estimation as we can only observe X_2 and cannot easily get at the form of $g(X_2)$. It is more often that biased/spurious results from misspecification (as in OLS-1) are obtained. Therefore, to reduce the impact of model misspecification, $\Phi_j(P)$ could be a simple and more robust solution to find the conditional linear association.

3.3 Asymptotic Performance

In this section, we study the asymptotic performance of the proposed estimator $\widehat{\Phi}_j$ in (3.10). Specifically, we first show the theoretical guarantees of $\widehat{\Phi}_j$ in general. Then, we consider the problem of estimating the regression coefficients in a linear model, and compare the performance of $\widehat{\Phi}_j$ with the OLS estimator when p is fixed, and with the de-biased lasso estimator when p is increasing with n, in terms of the limiting distribution.

3.3.1 Evaluating $\widehat{\Phi}_j - \Phi(P)$

Recall that the proposed estimator for the target nonparametric parameter $\Phi_j(P)$ is $\widehat{\Phi}_j = \frac{\widehat{\Psi}_j}{\widehat{V}_j}$ where $\widehat{\Psi}_j$ and $\widehat{V}_j$ are the estimators of $\Psi(P)$ and $V_j(P)$ defined in (3.9). By functional delta method, we know that the asymptotic properties of $\widehat{\Psi}_j$ are highly related to the properties of $\widehat{\Psi}_j$ and $\widehat{V}_j$. Chapter 2 have shown that,

$$\begin{aligned}\sqrt{n}\left(\widehat{\Psi}_j - \Psi_j(P)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{\Psi_j}(\mathbf{o}^{(i)}) + \Delta^{\Psi_j},\\ \sqrt{n}\left(\widehat{V}_j - V_j(P)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{V_j}(\mathbf{o}^{(i)}) + \Delta^{V_j}\end{aligned} \tag{3.12}$$

where $D_P^{\Psi_j}(\mathbf{o}^{(i)}) = \left[y^{(i)} - \mu_{P,Y}(\mathbf{x}_{-j}^{(i)}\right] \left[x_j^{(i)} - \mu_{P,j}(\mathbf{x}_{-j}^{(i)}\right]$ and $D_P^{V_j}(\mathbf{o}^{(i)}) = \left[x_j^{(i)} - \mu_{P,j}(\mathbf{x}_{-j}^{(i)}\right]^2$ are the non-parametric efficient influence functions of $\Psi_j(P)$ and $V_j(P)$ respectively. Δ^{Ψ_j} and Δ^{V_j} are the remainders. Hence, if we calculate the difference between $\widehat{\Phi}_j$ and $\Phi_j(P)$, we end up getting a very similar decomposition:

$$\begin{aligned}\sqrt{n}\left(\widehat{\Phi}_j - \Phi_j(P)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_P^{\Psi}(\mathbf{o}^{(i)}) - \Phi_j(P) \times D_P^{V}(\mathbf{o}^{(i)})}{V_j(P)} + \Delta^{\Phi_j} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{\Phi_j}(\mathbf{o}^{(i)}) + \Delta^{\Phi_j}.\end{aligned} \tag{3.13}$$

The following lemma states that the linear term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{\Phi_j}(\mathbf{o}^{(i)})$ is actually the canonical gradient (or equivalently the efficient influence function) of $\Phi_j(P)$. This analysis of the efficient influence function plays a critical role in establishing efficiency for asymptotically linear estimators in the context of nonparametric inference [14].

Lemma 3.1. *Given any $j \in \{1, ..., p\}$, the parameter $\Phi_j(P)$ is pathwise differentiable at $P \in \mathcal{P}$ and has efficient influence function $D_P^{\Phi_j}(\mathbf{o})$ of the form,*

$$D_P^{\Phi_j}(\mathbf{o}) := \frac{\{y - \mu_{P,Y}(\mathbf{x}_{-j})\}\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}}{V_j(P)} - \Phi_j(P)\frac{\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}^2}{V_j(P)}. \tag{3.14}$$

Under a linear model space $\mathcal{P}_{lm}$ where the realization $\mathbf{o} = (y, \mathbf{x})$ satisfies model (3.6), then this function (3.14) reduces to

$$D_{P_{lm}}^{\Phi_j}(\mathbf{o}) := \frac{\{x_j - \mu_{P_{lm},j}(\mathbf{x}_{-j})\}\epsilon}{V_j(P_{lm})}. \tag{3.15}$$

The proof is given in Appendix B.2. $\mu_{P_{\text{lm}},j}(\mathbf{x}_{-j})$ in (3.15) only refers to the linear relationship $E(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$, which does not imply linearity between X_j and $\mathbf{X}_{-j}$.

Lemma 3.1 gives the explicit form of the efficient influence function in linear first term of (3.13). However, to prove $\widehat{\Phi}_j$ is a consistent and efficient estimator, ones must show that the impact of Δ^{Φ_j} in (3.13) is asymptotically negligible as n increases. The following theorem

formally gives such a statement, and additionally provides the limiting distribution of $\widehat{\Phi}_j$.

Theorem 3.4. *Suppose that n observations $\{\mathbf{o}^{(i)}\}_{i=1}^n = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^n$ are i.i.d. generated from distribution P. Let $\widehat{\mu}_j$ and $\widehat{\mu}_Y$ be the estimators of $\mu_{P,j}$ and $\mu_{P,Y}$, having*

$$\begin{aligned} \int [\widehat{\mu}_Y(\mathbf{x}_{-j}) - \mu_{P,Y}(\mathbf{x}_{-j})]^2 dP(\mathbf{x}_{-j}) &= o_P\left(n^{-1/2}\right), \\ \int [\widehat{\mu}_j(\mathbf{x}_{-j}) - \mu_{P,j}(\mathbf{x}_{-j})]^2 dP(\mathbf{x}_{-j}) &= o_P\left(n^{-1/2}\right). \end{aligned} \tag{3.16}$$

Further assume that $\widehat{D}^{\Psi_j}(\mathbf{o})$ and $\widehat{D}^{V_j}(\mathbf{o})$ both fall in a P-Donsker class. Then, $\Delta^{\Phi_j} = o_P(1)$ and the plug-in estimator $\widehat{\Phi}_j$ is asymptotically normal*:*

$$\begin{aligned} \sqrt{n}\left[\widehat{\Phi}_j - \Phi_j(P)\right] &= \frac{1}{\sqrt{n}} \sum_{i=1}^n D_P^{\Phi_j}\left(\mathbf{o}^{(i)}\right) + o_P(1) \\ &\to N\left(0, \sigma_j^2(P)\right), \end{aligned} \tag{3.17}$$

where $\sigma_j^2(P) = \int \left\{D_P^{\Phi_j}(\mathbf{o})\right\}^2 dP(\mathbf{o})$.

The proof is given in Appendix B.3. The limiting distribution $\widehat{\Phi}_j$ is derived for a nonparametric space $\mathcal{P}$. That is, we do not assume any form of the relationship between the outcome Y and features $\mathbf{X}$. So, it is a very general result: As the sample size n increases, the centered estimator, i.e., $\widehat{\Phi}_j - \Phi_j(P)$, converges in distribution to a mean-zero normal variable with limiting variance $\sigma_j^2(P)$, under mild conditions.

As mentioned, $\Phi_j(P)$ reduces to the regression coefficient β_j when the linear relationship in (3.6) holds. It is then natural to ask, how does the estimator $\widehat{\Phi}_j$ compare to the classical OLS-based linear regression estimator? In next section, we shall focus on the scenario when linearity is present and compare $\widehat{\Phi}_j$ to the OLS estimator $\widehat{\beta}_j^{\text{OLS}}$ when p is fixed and to the de-biased lasso estimator $\widehat{\beta}_j^{\text{DL}}$ when p is increasing with n.

3.3.2 $\widehat{\Phi}_j$ versus $\widehat{\beta}_j^{OLS}$ when p is fixed

Suppose that n observations $\{\mathbf{o}^{(i)}\}_{i=1}^n = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^n$ are generated from the linear model (3.6) with fixed dimension p. In this setting, the standard tool to estimate β_j, for $j = 1, ..., p$, is the following OLS estimator:

$$\widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}, \tag{3.18}$$

where $\mathbf{x} \in \mathbb{R}^{n\times p}$ is the random design matrix where the ith row is the ith observation, i.e., $\mathbf{x} := [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}]^T$ and $\mathbf{y} = [y^{(1)}, ..., y^{(n)}]^T$. It is well-known that $\widehat{\beta}_j^{\text{OLS}}$ is asymptotically normal [77].

$$\sqrt{n}\left[\widehat{\beta}_j^{\text{OLS}} - \beta_j\right] \to_d N[0, \sigma^2\Theta_{p,jj}], \tag{3.19}$$

if $\mathbf{x}^{(i)} \sim_{i.i.d.} (\mathbf{0}, \Sigma_p)$, where Σ_p is invertible with $\Theta_p := \Sigma_p^{-1}$. $\Theta_{p,jj}$ is the (j, j)-th entry of Θ_p, for $j = 1, ..., p$. It has been shown that $\widehat{\beta}_j^{\text{OLS}}$ is the best linear unbiased estimator, which attains the Cramér–Rao bound [94].

Previously, we showed the equivalence of the parameters $\Phi_j(P)$ and β_j when linearity holds. Thus, $\widehat{\Phi}_j$ in (3.10) also estimates the regression coefficient β_j. The following theorem provides the asymptotic properties of $\widehat{\Phi}_j$ when linearity holds, arguing that its limiting distribution is the same as that of $\widehat{\beta}_j^{\text{OLS}}$.

Theorem 3.5. *Suppose that $\{\mathbf{o}^{(i)}\}_{i=1}^n = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^n$ are generated from linear model (3.6) with assumptions in Lemma 3.4 satisfied. Then, the estimator $\widehat{\Phi}_j$ defined in (3.10) is $\sqrt{n}$-consistent, and asymptotically normal, i.e.,*

$$\sqrt{n}\left[\widehat{\Phi}_j - \beta_j\right] \to_d N\left(0, \sigma_j^2(P_{lm})\right), \tag{3.20}$$

where $\sigma_j^2(P_{lm}) = \frac{\sigma^2}{V_j(P_{lm})}$. If we further suppose that $\mathbf{x}^{(i)} \sim_{i.i.d.} N(0, \Sigma_p)$ where Σ_p is invertible

with $\Theta_p := \Sigma_p^{-1}$, *then*

$$\sqrt{n}\left[\widehat{\Phi}_j - \beta_j\right] \to_d N\left(0, \sigma^2 \Theta_{p,jj}\right). \tag{3.21}$$

The proof is given in Appendix B.4. Theorem 3.5 implies that, using our proposed nonparametric parameter and implementing flexible techniques for estimating $\boldsymbol{\beta}$ does incur any additional first order costs (with a Gaussian design): The theoretical guarantees of model-based OLS estimators are maintained by using a more flexible and robust alternative $\widehat{\Phi}_j$ for estimating and making inference on the unknown regression coefficient β_j.

Another advantage of using $\widehat{\Phi}_j$ is that ones can avoid estimating two unknown quantities, σ^2 and $\Theta_{p,jj}$, appearing in the limiting distribution of $\widehat{\beta}_j^{\text{OLS}}$. Instead, to form a consistent estimator for the asymptotic variance $\sigma_j(P)$ of $\widehat{\Phi}_j$, two conditional means $\widehat{\mu}_Y$ and $\widehat{\mu}_j$, which have been obtained for $\widehat{\Phi}_j$, are sufficient. The estimator is of the following form:

$$\widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{D}^{\Phi_j}\left(\mathbf{o}^{(i)}\right)\right]^2, \tag{3.22}$$

which converges to its truth $\sigma_j^2(P) = \sigma^2\Theta_{p,jj}$ under the conditions in Theorem 3.5.

3.3.3 $\widehat{\Phi}_j$ versus $\widehat{\beta}_j^{DL}$ when $p > n$ and $p, n \to \infty$

Previous results are discussed under the setting when p is fixed. However, in the high-dimensional setting, i.e., $p > n$, things get more complicated as: (i) the estimated covariance matrix $\mathbf{x}^T\mathbf{x}$ is singular, so the OLS method collapses; (ii) some regularity conditions under fixed p may fail to hold as p is increasing as well. Extensive work has been done developing methods and theory for the estimation of regression coefficients in the high-dimensional regime. Most of the theory is non-asymptotic and gives finite sample concentration inequalities. There has also been substantial work on inference in high-dimensional problems, however some of the asymptotic arguments have been a bit informal. In particular, almost none of the work we have seen formally defines an asymptotic regime where n and p both grow.

We formalize these ideas by using a triangular array of random variables to specify an

asymptotic regime where p can change with n. Here, we formally compare the asymptotic properties of de-biased lasso estimator $\widehat{\beta}_j^{\mathrm{DL}}$ and our proposed estimator $\widehat{\Phi}_j$ with respect to such a triangular array of observations, in the high-dimensional setting.

Setup for triangular array. Suppose that the random variables of covariates $\mathbf{X}_n = (X_{n,1}, ..., X_{n,p_n})^T$ have the following triangular array structure:

$$\begin{aligned} &X_{1,1}, ..., X_{1,p_1} \\ &X_{2,1}, ..., X_{2,p_2} \\ &\vdots \\ &X_{n,1}, ..., X_{n,p_n}. \end{aligned} \tag{3.23}$$

In this setup, the generated data $\{\mathbf{o}_n^{(i)}\}_{i=1}^n = \{(y_n^{(i)}, \mathbf{x}_n^{(i)})\}_{i=1}^n$ with $\mathbf{x}_n^{(i)} = (x_{n,1}^{(i)}, ..., x_{n,p_n}^{(i)})^T$ have joint distribution which we denote P_n: Observations $\mathbf{x}_n^{(1)}, \ldots, \mathbf{x}_n^{(n)}$ in the same row are independently and identically distributed, while those in different rows are allowed to have different distributions. To make our limiting statement consistent with this triangular array setup, we say a random variable $X_n = o_{P_n}(a_n)$ if $\lim_{n\to\infty} P_n(|X_n/a_n| \geq \tau) = 0$, $\forall \tau > 0$.

For a fixed sample size n, we assume that the data satisfy the following linear relationship

$$Y_n = \mathbf{X}_n^T \boldsymbol{\beta}_n + \epsilon_n = \sum_{j=1}^{p_n} \beta_{n,j} X_{n,j} + \epsilon_n, \tag{3.24}$$

for $j = 1, ..., p_n$ and $n = 1, 2, ...$ The error term has $\epsilon_n \sim N(0, \sigma^2) \perp\!\!\!\perp X_{n,j}$, The dimension $p = p_n$, the joint distribution of the features $\mathbf{x}_n$, and the distribution of the outcome Y_n are now allowed to be dependent on the sample size n.

To estimate $\beta_{n,j}$ when $p_n > n$, [111] proposed the de-biased lasso, to overcome the issue of bias with the standard lasso estimator and account for the uncertainty of the estimator. In that work, a number of assumptions are required for asymptotically correct inference. Here we transport and formalize those assumptions for our triangular array setup.

Assumption 3.2. *Suppose that n observations $\left\{(y_n^{(i)}, \mathbf{x}_n^{(i)})\right\}_{i=1}^n$ are generated from model*

(3.24), *with the following assumptions satisfied:*

1. $\mathbf{x}_n^{(i)} \in \mathbb{R}^{p_n} \sim_{i.i.d.} N(0, \Sigma_{p_n})$ *for* $i = 1, 2, \ldots, n$.

2. $\Sigma_{p_n} := \mathrm{E}\left[(\mathbf{x}_n^{(i)})^T \mathbf{x}_n^{(i)}\right] \in \mathbb{R}^{p_n \times p_n}$ *is invertible with* $\Theta_{p_n} := \Sigma_{p_n}^{-1}$.

3. $\|\Sigma_{p_n}\|_\infty = \mathcal{O}(1)$.

4. *The minimum eigenvalue* $\Lambda^2_{\min}(\Sigma_{p_n})$ *is uniformly bounded away from 0.*

With these assumptions, we are able to characterize the asymptotic behaviour of the de-biased lasso estimator, which is calculated by

$$\widehat{\boldsymbol{\beta}}_n^{DL} = \widehat{\boldsymbol{\beta}}_n^{Lasso} + \widehat{\Theta}_{p_n} \mathbf{x}_n^T \left(\mathbf{y}_n - \mathbf{x}_n \widehat{\boldsymbol{\beta}}_n^{Lasso}\right)/n, \tag{3.25}$$

where

$$\widehat{\boldsymbol{\beta}}_n^{Lasso} = \underset{\boldsymbol{\beta}_n \in \mathbb{R}^{p_n}}{\operatorname{argmin}} \{\|\mathbf{y}_n - \mathbf{x}_n \boldsymbol{\beta}_n\|_2^2/n + 2\lambda \|\boldsymbol{\beta}_n\|_1\}. \tag{3.26}$$

$\widehat{\Theta}_{p_n}$ is an estimator of Θ_{p_n} by node-wide lasso regression [75]. That is, for each $j = 1, ..., p_n$, we estimate

$$\widehat{\boldsymbol{\gamma}}_{n,j} := \underset{\boldsymbol{\gamma}_n \in \mathbb{R}^{p_n - 1}}{\operatorname{argmin}} \left(\|\mathbf{x}_{n,j} - \mathbf{x}_{n,-j} \boldsymbol{\gamma}_n\|_2^2/n + 2\lambda_j \|\boldsymbol{\gamma}_n\|_1\right), \tag{3.27}$$

where $\mathbf{x}_{n,j}$ is the jth column of the design matrix $\mathbf{x}_n$ and $\widehat{\boldsymbol{\gamma}}_{n,j} = \{\widehat{\gamma}_{n,jk} : k = 1, ..., p_n, k \neq j\}$. Then the node-wise lasso estimator of Θ_{p_n} is given by

$$\widehat{\Theta}_{p_n} := \widehat{\Gamma}_n^{-2} \widehat{\mathcal{C}}_n, \tag{3.28}$$

where $\widehat{\Gamma}_n := \mathrm{diag}(\widehat{\tau}^2_{n,1}, ..., \widehat{\tau}^2_{n,p_n})$ with $\widehat{\tau}^2_{n,j} := \|\mathbf{x}_{n,j} - \mathbf{x}_{n,-j} \widehat{\boldsymbol{\gamma}}_{n,j}\|_2^2/n + \lambda_j \|\widehat{\boldsymbol{\gamma}}_{n,j}\|_1$. $\widehat{\mathcal{C}}_n^2$ is defined

as

$$\widehat{\mathcal{C}}_n^2 := \begin{pmatrix} 1 & -\widehat{\gamma}_{n,21} & \cdots & -\widehat{\gamma}_{n,p_n1} \\ -\widehat{\gamma}_{n,12} & 1 & \cdots & -\widehat{\gamma}_{n,p_n2} \\ \vdots & \ddots & \vdots & \vdots \\ -\widehat{\gamma}_{n,1p_n} & -\widehat{\gamma}_{n,2p_n} & \cdots & 1 \end{pmatrix}_{p_n \times p_n}. \tag{3.29}$$

So, the jth column of $\widehat{\Theta}_{p_n}$ is $\widehat{\Theta}_{p_n,j} := \widehat{\mathcal{C}}_{n,j}/\widehat{\tau}_{n,j}^2$, where $\widehat{\mathcal{C}}_{n,j}$ is the jth columns of $\widehat{\mathcal{C}}_n$.

The form of (3.25) implies that the asymptotic performance of $\widehat{\boldsymbol{\beta}}_n^{\mathrm{DL}}$ will depend on how well Θ_{p_n} is estimated. Therefore, to ensure that node-wise lasso algorithm can work well, ones need to specify additional sparsity assumptions with respect to the columns of Θ_{p_n}, i.e.,

$$s_{n,j} := |\{k \neq j : \Theta_{p_n,jk} \neq 0\}|. \tag{3.30}$$

The authors in [111] obtain the asymptotic behavior of heuristically $\widehat{\beta}_{n,j}^{\mathrm{DL}}$ by showing $\sqrt{n}\left(\widehat{\beta}_{n,j}^{\mathrm{DL}} - \beta_{n,j}\right) = \Theta_{p_n,j}^T \mathbf{x}_n^T \boldsymbol{\epsilon_n}/\sqrt{n} + o_P(1)$, and conditioning on $\mathbf{x}_n$. Thus, with Gaussian errors one has that $\Theta_{p_n,j}^T \mathbf{x}_n^T \boldsymbol{\epsilon_n}/\sqrt{n}|\mathbf{x}_n \sim N(0, \sigma^2 \Theta_{p_n,jj})$. However, they did not explicitly engage with the asymptotic behavior when the dimension grows. Here we formally provide such asymptotic results for the de-biased lasso estimator $\widehat{\boldsymbol{\beta}}_{n,j}^{\mathrm{DL}}$, using a triangular array argument.

Theorem 3.6. *Suppose n observations $\{(y_n^{(i)}, \mathbf{x}_n^{(i)})\}_{i=1}^n$ are generated from linear model (3.24), with Assumption 3.2 satisfied. Assume that the model is sparse $s_{n,0} := |1 \leq j \leq p_n : \beta_{n,j} \neq 0| = o(\sqrt{n}/\log p_n)$, and the columns sparsity of Θ_{p_n} satisfies $\max_{1 \leq j \leq p_n} s_{n,j} = o(n/\log p_n)$. If the regularization parameters λ in (3.26) and λ_j in (3.28) are suitably selected with $\lambda \asymp \sqrt{\log p_n/n}$ and $\lambda_j \asymp \sqrt{\log p_n/n}$, $\forall j = 1, ..., p_n$. The de-biased lasso estimator $\widehat{\beta}_{n,j}^{DL}$ defined in (3.25) has*

$$\frac{\sqrt{n}\left(\widehat{\beta}_{n,j}^{DL} - \beta_{n,j}\right)}{\sigma\sqrt{\Theta_{p_n,jj}}} = R_n + o_{P_n}(1), \tag{3.31}$$

where $R_n = (r_n^{(1)} + r_n^{(2)} + ... + r_n^{(n)})/\sqrt{n}$ with $r_n^{(i)} = \frac{\Theta_{p_n,j}^T \mathbf{x}_n^{(i)} \epsilon^{(i)}}{\sigma\sqrt{\Theta_{p_n,jj}}}$. Moreover, $r_n^{(i)} \equiv r_n$ with

$\mathrm{E}(r_n) = 0$ *and* $\mathrm{Var}(r_n) = 1$. *If* $\mathrm{E}\,|r_n^3| = O(1)$, *then*

$$\mathbb{P}\left(R_n \geq z\right) - \mathbb{P}(Z > z) = O\left[\left(\mathrm{E}\,|r_n|^3/\sqrt{n}\right)^{1/4}\right] = o(1), \tag{3.32}$$

where Z *is a standard normal variable.*

The details of the proof are provided in Appendix B.7. Theorem 3.6 roughly says that, as long as the number of features does not grow too quickly then the remainder term $\Delta_{n,j}^{\mathrm{DL}}$ does not contribute to the first order behaviour of the debiased lasso estimate. Thus, the limiting distribution of $\widehat{\beta}_{n,j}^{\mathrm{DL}}$ is dominated by the linear term $\Theta_{p_{n,j}}^T \mathbf{x}_n^T \boldsymbol{\epsilon}$. After scaling by its limiting variance, the de-biased lasso estimator would converge in distribution to a standard normal variable.

In this setting, we can also use $\widehat{\Phi}_j$ when $p = p_n > n$ to estimate $\beta_{n,j}$. Here, we reiterate our results of $\widehat{\Phi}_j$ in Section 3.3.1 using the triangular array of random variables.

Proposition 3.3. *Suppose that* n *observations* $\{\mathbf{o}_n^{(i)}\}_{i=1}^n = \{(y_n^{(i)}, \mathbf{x}_n^{(i)})\}_{i=1}^n$ *have joint distribution* P_n. *Then, the estimator for the target parameter* $\Phi_j(P_n)$ *is given by*

$$\widehat{\Phi}_{n,j} := \frac{\widehat{\Psi}_{n,j}}{\widehat{V}_{n,j}} = \frac{\frac{1}{n}\sum_{i=1}^n \left[y_n^{(i)} - \widehat{\mu}_{n,Y}(\mathbf{x}_{n,-j}^{(i)})\right]\left[x_{n,j}^{(i)} - \widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}^{(i)})\right]}{\frac{1}{n}\sum_{i=1}^n \left[x_{n,j}^{(i)} - \widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}^{(i)})\right]^2}, \tag{3.33}$$

where $\widehat{\mu}_{n,Y}$ *and* $\widehat{\mu}_{n,j}$ *are the estimated conditional means of* $\mu_{P_n,Y}$ *and* $\mu_{P_n,j}$. *The difference between* $\widehat{\Phi}_{n,j}$ *and* $\Phi_j(P_n)$ *satisfies*

$$\sqrt{n}\left(\widehat{\Phi}_{n,j} - \Phi_j(P_n)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n^{(i)}) + \Delta_n^{\Phi_j}. \tag{3.34}$$

where $D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n) = \frac{\{y_n - \mu_{P_n,Y}(\mathbf{x}_{n,-j})\}\{x_{n,j} - \mu_{P_n,j}(\mathbf{x}_{n,-j})\}}{V_j(P_n)} - \Phi_j(P_n)\frac{\{x_j n,j - \mu_{P_n,j}(\mathbf{x}_{n,-j})\}^2}{V_j(P_n)}$ *is the efficient influence function of* $\Phi_j(P_n)$. *In particular, when* $\left\{\mathbf{o}_n^{(i)}\right\}_{i=1}^n$ *are generated from linear model (3.24), then*

$$\Phi_j(P_{n,lm}) = \beta_{n,j}, \tag{3.35}$$

with efficient influence function $D^{\Phi_{n,j}}_{P_{n,lm}}(\mathbf{o}_n) = \frac{[x_{n,j} - \mu_{P_n,j}(\mathbf{x}_{n,-j})]\epsilon_n}{V_j(P_{n,lm})}$.

Similar to the low-dimensional case, the asymptotic performance of $\widehat{\Phi}_{n,j}$ will be dominated by the linear first term in (3.34) if $\Delta_n^{\Phi_j}$ converges to zero sufficiently fast, i.e., if $\Delta_n^{\Phi_j} = o_{P_n}(1)$.

Lemma 3.2. *Suppose that* n *observations* $\{\mathbf{o}_n^{(i)}\} = \{(y_n^{(i)}, \mathbf{x}_n^{(i)})\}_{i=1}^n$ *are generated from linear model* (3.24). *Let* $\widehat{\mu}_{n,j}$ *and* $\widehat{\mu}_{n,Y}$ *be the estimator the conditional means* $\mu_{P_n,j}$ *and* $\mu_{P_n,Y}$, *having*

$$\begin{aligned} \int [\widehat{\mu}_{n,Y}(\mathbf{x}_{n,-j}) - \mu_{P_n,Y}(\mathbf{x}_{n,-j})]^2 dP_n(\mathbf{x}_{n,-j}) &= o_{P_n}\left(n^{-1/2}\right), \\ \int [\widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}) - \mu_{P_n,j}(\mathbf{x}_{n,-j})]^2 dP_n(\mathbf{x}_{n,-j}) &= o_{P_n}\left(n^{-1/2}\right). \end{aligned} \tag{3.36}$$

Further assume that

$$\begin{aligned} \sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}_n^{\Psi_j}(\mathbf{o}_n) - D_{P_n}^{\Psi_j}(\mathbf{o}_n)\right] &= o_{P_n}(1), \\ \sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}_n^{V_j}(\mathbf{o}_n) - D_{P_n}^{V_j}(\mathbf{o})\right] &= o_{P_n}(1). \end{aligned} \tag{3.37}$$

Then, the remainder $\Delta_n^{\Phi_j}$ *can be bounded*

$$\Delta_n^{\Phi_j} = o_{P_n}(1). \tag{3.38}$$

Remark. An illustrative example of the conditional mean estimator in the high-dimensional case is the lasso estimator. Let $\widehat{\mu}_{n,j}$ and $\widehat{\mu}_{n,Y}$ be the lasso estimates of $\mu_{P_n,j}$ and $\mu_{P_n,Y}$. If the linear model (3.24) is sparse with $s_0 = o\left(\sqrt{n/\log p_n}\right)$, then (3.36) will hold when choosing the regularization parameter properly, i.e., $\lambda \asymp \sqrt{\log p_n/n}$. This assumption on the sparsity of the model is less restrictive compared to de-biased lasso as in Theorem 3.6, which requires $s_0 = o\left(\sqrt{n}/\log p_n\right)$.

Lemma 3.2 shows that if (i) the conditional means converge sufficiently quickly and (ii) the empirical process term in (3.37) can be bounded in triangular array setup, then $\Delta_n^{\Phi_j}$ will not have a first order contribution to the asymptotic behavior of our estimator. Thus, we give the following main result for the asymptotic performance of $\widehat{\Phi}_{n,j}$ in the high-dimensional

setting.

Theorem 3.7. *Suppose that* $\{\mathbf{o}_n^{(i)}\} = \{(y_n^{(i)}, \mathbf{x}_n^{(i)})\}_{i=1}^n$ *are generated from distribution* P_n *with assumptions in Lemma 3.2 satisfied. Then, the estimator* $\widehat{\Phi}_{n,j}$ *in (3.33),* $(j = 1, ..., p_n)$*, have*

$$\frac{\sqrt{n}(\widehat{\Phi}_{n,j} - \beta_{n,j})}{\sigma_j(P_n)} = W_n + o_{P_n}(1), \tag{3.39}$$

where $W_n = (w_n^{(1)} + w_n^{(2)} + \cdots + w_n^{(n)})/\sqrt{n}$ *with* $w_n^{(i)} = \frac{D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n^{(i)})}{\sigma_j(P_n)}$*.* $w_n^{(i)} \equiv w_n$ *with* $\mathrm{E}(w_n) = 0$ *and* $\mathrm{Var}(w_n) = 1$*. If* $E|w_n|^3 = O(1)$*, then we have*

$$\mathbb{P}(W_n \leq z) - \mathbb{P}(Z \leq z) = O\left[\left(\mathrm{E}\,|w_n|^3\sqrt{n}\right)^{1/4}\right] = o(1), \tag{3.40}$$

where Z *is a standard normal variable. If we further suppose that data satisfy the linear model (3.24) with* $\mathbf{x}_n^{(i)} \sim_{i.i.d.} N(0, \Sigma_{p_n})$*, then*

$$\frac{\sqrt{n}(\widehat{\Phi}_{n,j} - \beta_{n,j})}{\sigma\Theta_{p_n,jj}} = W_n + o_{P_n}(1), \tag{3.41}$$

with $w_n^{(i)} = \frac{[x_{n,j} - \mu_{P_n,j}(\mathbf{x}_{n,-j})]\epsilon_n}{\sigma^2}$ *and (3.40) still holds.*

The details of the proof are provided in Appendix B.6. Therefore, in the high-dimensional setting, the proposed estimator is still asymptotically consistent and normal. In particular, under the presence of linearity and normality of the covariates, $\widehat{\Phi}_{n,j}$ has the same limiting distribution as the de-biased lasso estimator $\widehat{\beta}_{n,j}^{\mathrm{DL}}$. However, unlike $\widehat{\beta}_{n,j}^{\mathrm{DL}}$ which demands additional assumptions of the data and model, e.g., Assumption 3.2 and some sparsity conditions, the asymptotic guarantees of $\widehat{\Phi}_{n,j}$ only depend on the convergence of the estimated conditional means: The above results remain valid even, eg., under a partially linear model.

3.4 Simulation study

We assess the performance of the proposed (theoretically optimal) estimators $\widehat{\Phi}_j$ in (3.10) for the conditionally linear association measure $\Phi_j(P)$. Specifically, we compare $\widehat{\Phi}_j$ with the

de-biased lasso estimator for estimating the regression coefficients in high-dimensional linear models. The performance will be empirically evaluated by (i) bias, (ii) variance, and (iii) the coverage of nominal 95% Wald-type confidence intervals.

3.4.1 When features are uncorrelated

We consider generating data from the following linear model:

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \ldots + X_{255}\beta_{255} + \epsilon, \quad \epsilon \sim N(0,1) \perp\!\!\!\perp X \\ X &\sim N(\mathbf{0}, I_{255}), \quad \beta = (\beta_1, \ldots, \beta_{255})^T = (\underbrace{1, \ldots, 1}_{5}, \underbrace{0, \ldots, 0}_{250})^T. \end{aligned} \tag{3.42}$$

We generate random datasets of size $n \in \{100, 500, 1000, 2000, 3000\}$ and estimate β_1 and β_6 in each case, where the true values are respectively 1 and 0. For each size, we run 100 simulations. We estimate $\beta_j, j = 1, 6$ by lasso $\widehat{\beta}_j^{\text{Lasso}}$ in (3.26), de-biased lasso $\widehat{\beta}_j^{\text{DL}}$ in (3.25), and the proposed estimator $\widehat{\Phi}_j$ in (3.10) respectively. For $\widehat{\beta}_j^{\text{Lasso}}$ and $\widehat{\beta}_j^{\text{DL}}$, the regularization parameters are separately tuned by cross validation. As for $\widehat{\Phi}_j$, the conditional means are estimated by the lasso.

The results for estimating $\beta_1 = 1$ and $\beta_6 = 0$ are provided in Figure 3.2, in the bottom and top panel respectively. It shows that the empirical $\sqrt{n}$-scaled bias of our theoretically optimal plug-in estimator $\widehat{\Phi}_j$ and de-biased lasso estimator $\widehat{\beta}_j^{\text{DL}}$ vanish to 0 as n goes to infinity for both β_1 and β_6. However, this is not the case for lasso estimator: Its $\sqrt{n}$-scaled bias appears to be roughly constant when the true coefficient is non-zero. This is generally well known and is the rationale for using the debiased lasso. The empirical variances of $\widehat{\Phi}_j$ and $\widehat{\beta}_j^{\text{DL}}$ stabilize when scaled by n. In particular, they are very close to each other, which agrees with the results in Theorem 3.7. As expected, the confidence intervals of $\widehat{\Phi}_j$ and $\widehat{\beta}_j^{\text{DL}}$ also converge to the nominal 95% as n increases. We do not provide the bootstrapped-based interval for lasso estimators, as in Chapter 2 we have shown its failure.

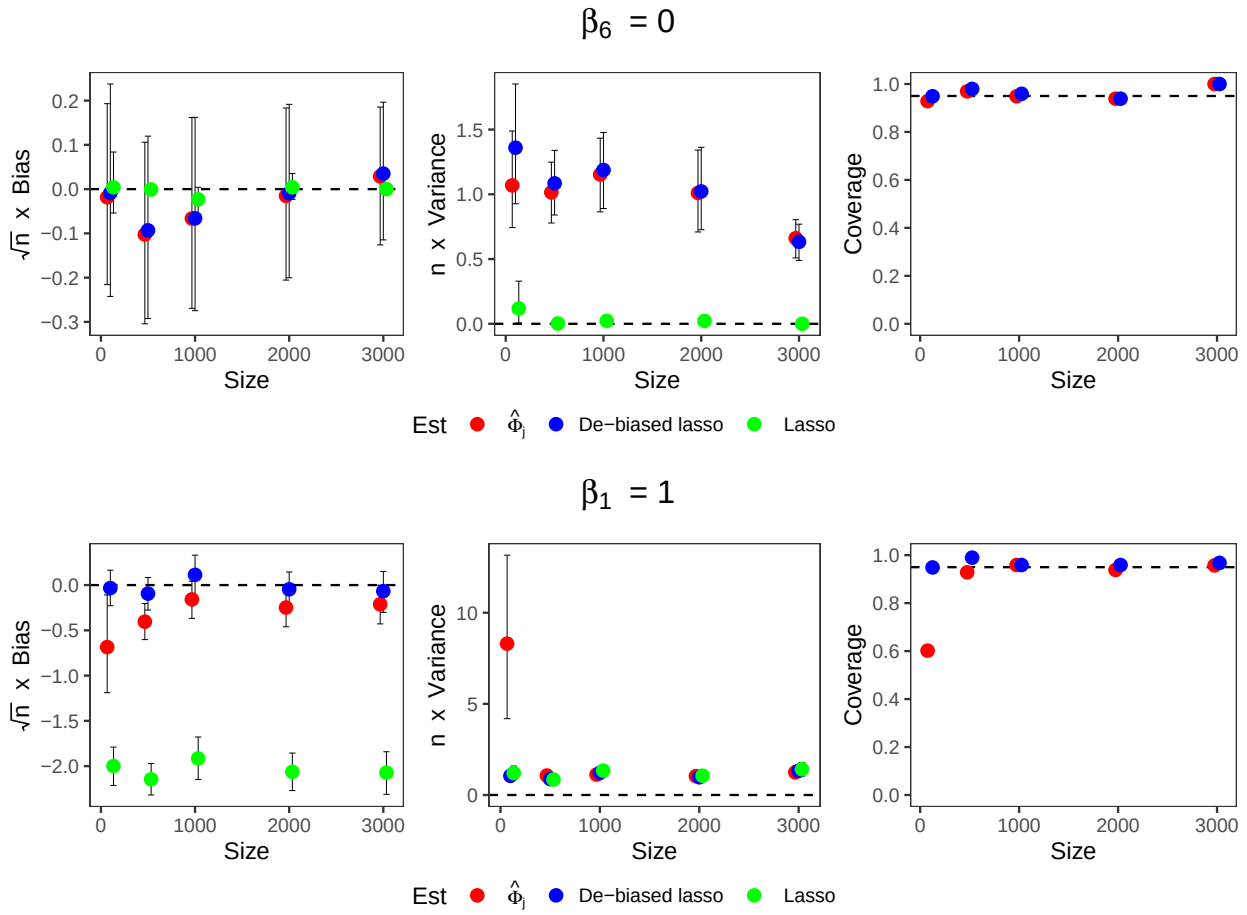

Figure 3.2: When features are uncorrelated: Empirical $\sqrt{n}-$scaled bias, Empirical n-scaled variance and empirical coverage of 95% confidence interval of $\widehat{\Phi}_j$ (red), $\widehat{\beta}_j^{\text{DL}}$ (blue), and $\widehat{\beta}_j^{\text{Lasso}}$ (green). The results for estimating $\beta_1 = 1$ and $\beta_6 = 0$ are respectively provided in the top and bottom panel.

3.4.2 When features are correlated

Now, we still specify the same relationship between Y and $\mathbf{X}$ as in (3.42) but add correlation to features. That is, $X \sim N(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Sigma_{1\sim5} & 0 & \cdots & 0 \\ 0 & \Sigma_{6\sim15} & \cdots & 0 \\ \cdots & \cdots & \ddots & \\ 0 & 0 & 0 & \Sigma_{246\sim255} \end{pmatrix}. \tag{3.43}$$

For each of $\Sigma_{1\sim5}$, ..., $\Sigma_{246\sim255}$, we use a symmetric matrix with diagonal elements of 1 and off-diagonal elements of 0.3.

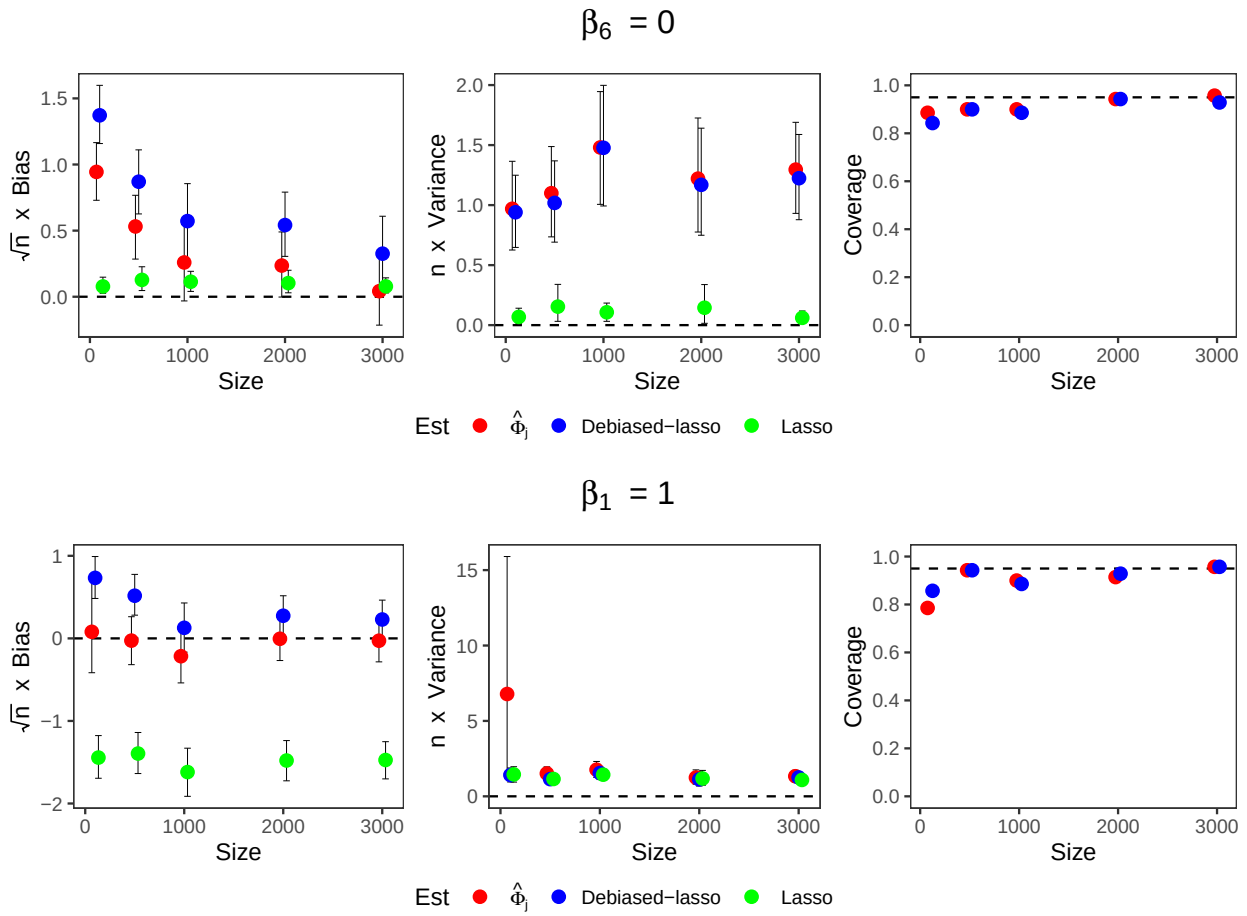

Figure 3.3: When features are correlated: Empirical $\sqrt{n}-$scaled bias, Empirical n-scaled variance and empirical coverage of 95% confidence interval of $\widehat{\Phi}_j$ (red), $\widehat{\beta}_j^{\mathrm{DL}}$ (blue), and $\widehat{\beta}_j^{\mathrm{Lasso}}$ (green). The results for estimating $\beta_1 = 1$ and $\beta_6 = 0$ are respectively provided in the top and bottom panel.

The results are shown in Figure 3.3. When features are correlated, we see that the de-biased lasso estimator $\widehat{\beta}_j^{\mathrm{DL}}$ becomes less unbiased regardless of whether the true coefficient to be estimated is zero or non-zero. However, our proposed $\widehat{\Phi}_j$ remains effective. Other patterns are similar to what we observed in Figure 3.2.

3.5 Discussion

In this chapter, we introduce a nonparametric measure to assess the conditional linear association between the outcome and any explanatory variables of interests. Traditionally, a parametric (linear) model is used to address this type of problem. Here, we propose parameter that is model-agnostic which can reduce to (i) the regression coefficients when linearity (or partial linearity) holds and (ii) the variance weighted treatment effect when the covariate of interest is binary. We additionally propose a simple and natural plug-in estimator of this parameter, which is asymptotically normal and non-parametrically efficient: These theoretical guarantees always holds even without linearity. Additionally, if the outcome and covariates truly have a linear relationship, and the covariates have a joint Gaussian distribution then our estimator has the same limiting distribution as OLS estimator when data is low-dimensional and with de-biased lasso estimator when data is high-dimensional.

One major strength of this work is in that the proposed estimator can be generally effective in estimating the target parameter $\Phi_j(P)$. The efficacy of the estimator relies almost entirely on our ability to build good predictive models for estimating two conditional means. In estimating those conditional means we can leverage machine learning techniques most appropriate for our specific problem. This is very different from model-based techniques for linear association analysis, like OLS method and the de-biased lasso. Their efficacy and asymptotic properties highly depend on the correctness of those parametric specification. This also increases the flexibility of our method: With no modification, it allows for inference of a regression coefficient under, eg., partial linearity. In addition, in high-dimensional scenarios, it can be difficult to make inference even in the presence of a linear model: Methods like the debiased lasso require estimation of quantities like the precision matrix and variance of the error to form confidence intervals [134, 126]. In contrast, for our proposal, estimating two conditional means well is sufficient for estimating our parameter and subsequent inference.

One potential concern with our method is substantial computational cost in high throughput screening problems. We require a pair of estimated regressions for every feature that

we would like to consider as a primary exposure. Something like the graphical lasso [75], might be more computationally efficient in such scenarios (though in cases where the features are not jointly Gaussian, the scientific relevance of such estimates is unclear). In addition, one would need to address multiple comparisons in such scenarios: Strategies like Bonferroni correction [80] and Benjamini-Hochberg Procedure [11] can be used.

Chapter 4

ON THE OPTIMALITY OF NUCLEAR-NORM-BASED MATRIX COMPLETION FOR PROBLEMS WITH SMOOTH NON-LINEAR STRUCTURE

4.1 Introduction

Matrix completion is a framework that has gained popularity in a wide range of machine learning applications, including recommender systems [63], system identification [72], global positioning [97] and natural language processing [122]. It is a useful framework for complex prediction problems, where each observation comes with a heterogeneous collection of observed features. In particular, matrix completion is applied to problems where the object of inference or prediction is a matrix whose rows correspond to observation and columns to variables/features. In many applications, only a subset of entries in this matrix are observed (often with noise), and the goal is to "complete" the matrix, filling in estimates of the unobserved entries. This "completion" is done by leveraging the known structure in the matrix. The most famous example, which brought matrix completion to prominence, is the Netflix Challenge [63], where a small sample of observed ratings for each customer was used to successfully predict future/unobserved movie ratings for Netflix customers.

More formally, suppose we have an underlying unobserved matrix $M \in \mathbb{R}^{n\times p}$: We then observe a subset of the entries from the noise-contaminated matrix $Y = M + E$, where E is a matrix of i.i.d. mean zero, finite variance noise variables. Our goal is to recover matrix M from this partially observed, noisy Y. This is known as matrix completion. Without any structure on the matrix M, recovering the values of M corresponding to unobserved entries is impossible [66]. Matrix completion becomes possible if one imposes some constraints on the structure of the underlying matrix: It is most common to assume that M is low

rank. Directly employing this assumption by e.g., finding the minimum rank completion of Y (or corresponding rank-constrained regression) is unfortunately NP-hard and becomes computationally infeasible for problems involving large matrices [21, 26]. Over the last decades, computationally efficient methods using convex optimization have been developed for recovering a low rank matrix from a small number of observations with near-optimal statistical guarantees in primarily noiseless problems [99, 84, 21, 85], and when the observed entries are contaminated with noise [20, 61]. These methods rely on using the nuclear norm of the matrix [40, 56], i.e., sum of its singular values, as a convex surrogate for the matrix rank. The low-rank structure leveraged in matrix completion can be thought of as learning a linear embedding of the data in a low-dimensional space.

In practice, the underlying matrix M may not be low rank. However, we often believe it may still have useful low-dimensional structure. It has thus become popular to learn a low-dimensional non-linear embedding of the data. This idea is used both in matrix completion and more generally for low-dimensional summaries of data. It has been applied in motion recovery [124], epigenomics [93], and health data analytics [118] among other areas. To recover these embeddings, Reproducing Kernel Hilbert Space (RKHS) methods [38], nearest neighbor methods [67], and deep learning methods like autoencoders and neural-network-based variational frameworks [37, 128, 57] have been used.

Additionally, there has been strong empirical evidence that matrix completion methods based on nuclear norm penalization perform well even in scenarios where any low dimensional structure is likely non-linear. As these methods were developed for linear low rank structure, this is, at first glance, a bit surprising. There has been some work giving theoretical justification for these empirical results [24, 108]. In particular, they note that in the presence of some types of non-linear low-dimensional structure in M, nuclear norm-based matrix completion methods can still consistently estimate M. These work additionally gives some non-stochastic approximation error results. However, optimality of the statistical perform of nuclear-norm-based matrix completion is not considered to the best of our knowledge.

In this manuscript, we delve further into the performance of matrix completion for M

with low-dimensional, non-linear structure. In particular, we consider M with rows that can be embedded in a low-dimensional smooth manifold. We then (i) show that nuclear norm-based matrix completion can consistently estimate M; (ii) characterize the rate at which the reconstruction error converges to 0 as a function of the size of the matrix, number of observed entries, and smoothness and dimension of the underlying manifold; and (iii) prove that, up to a log term, this rate cannot by improved upon by any method; that is, our upper bound is actually the minimax rate optimal for reconstruction error in this problem. Furthermore, our error bounds (and our techniques) also relate the matrix completion problem clearly to more classical non-parametric estimation: Our reconstruction error bounds parallel the minimax rate of mean squared error (MSE) in the nonparametric regression setting. Results (ii) and (iii), we believe, are novel.

Our experiments on synthetic data corroborate our theoretical findings. In particular, they suggest that the finite sample empirical performance of matrix completion in non-linear low rank embeddings is consistent with the asymptotic theoretical error bounds. These empirical results also corroborate the claim that better performance is achieved when the embedding of the underlying matrix M lies in a smoother manifold.

4.2 Methods

4.2.1 Problem setup

We start by giving some notation. We use upper case letters to represent matrices and lower case letters to represent scalars. The trace inner product of any two matrices, $M, B \in \mathbb{R}^{n\times p},\ n, p \in \mathbb{Z}^+$, is $\langle M, B\rangle = \mathrm{tr}(M^T B)$. The element-wise infinity norm of $M \in \mathbb{R}^{n\times p}$ is defined by $\|M\|_\infty = \max_{1\le i\le n, 1\le j\le p} |m_{ij}|$ where m_{ij} denotes the (i, j)-th entry of M. We also denote the Frobenius norm of matrix M as $\|M\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p m_{ij}^2}$.

In the general matrix completion problem, we randomly observe some of the entries from a matrix $M \in \mathbb{R}^{n\times p}$; the observed entries may also be contaminated with error. To support our later theoretical derivations, we will describe this process in terms of a set of mask matrices

$X_t \in \mathbb{R}^{n\times p}$ and observed values $y_t \in \mathbb{R}$. Each X_t is a matrix with a single 1 whose position is indexed by t and all other entries are equal to 0 as follows:

$$X_t = \begin{pmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}_{n\times p} . \tag{4.1}$$

The collection of matrices X_t fall in the set $\mathcal{X} = \{e_n(i)e_p(j)^T, \text{ for all } i = 1,\ldots,n \text{ and } j = 1,\ldots,p\}$, where $e_n(i) \in \mathbb{R}^n$ is the basis vector consisting of all zeros except for a single 1 at ith entry. In this formulation, X_t indicates the location in M where y_t is drawn from. That is, for $X_t = e_n(i)e_p(j)^T \in \mathcal{X}$, $\langle X_t, M\rangle = m_{ij}$.

Now, we can frame the matrix completion problem as follows: Suppose we have N pairs of observations (X_t, y_t), $t = 1,\ldots,N$, that satisfy

$$y_t = \langle X_t, M\rangle + \xi_t, \tag{4.2}$$

where ξ_t are i.i.d random errors distributed $N(0,\sigma^2)$, $M \in \mathbb{R}^{n\times p}$ is the underlying true matrix to be recovered, and $y_t \in \mathbb{R}$ are observed values. The observed matrix can be written as $Y = \sum_{t=1}^{N} y_t X_t$ where N is the number of observed entries. We assume that X_t is uniformly sampled at random (USR) from $\mathcal{X}$ [61], i.e. $X_t \sim \Pi$, and the probability that the (i,j)th entry of X_t equals to 1 is $\pi_{ij} = \mathrm{P}(X_t = e_i(n)e_j(p)^T) = \frac{1}{np}$ for $1 \le i \le n, 1 \le j \le p$. This is essentially a missing completely at random (MCAR) assumption.

The goal is to recover M given pairs (X_t, y_t), $t = 1, 2, ..., N$, and we are generally interested in the setting where $N \ll np$. To solve this problem, existing methods often assume that M has low rank (or approximately low rank), i.e. $M \simeq UV^T$ with $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{p\times r}$ for some integer $r \ll \min(n,p)$. In contrast to this low rank assumption, this paper studies the problem where M is not necessarily low-rank but generated from a low-dimensional non-linear

manifold. This notion is formalized in the next section.

4.2.2 Non-linearly Embeddable Matrices

We begin by formalizing what we mean by "low-dimensional non-linear structure". Consider a matrix M, a positive integer K, and a function class $\mathcal{F} \subset \mathcal{L}^2\left(\mathbb{R}^K\right)$. We say M is $\mathcal{F}$-embeddable if there exist functions $f_j \in \mathcal{F} : \mathbb{R}^K \to \mathbb{R},\ j = 1, \ldots, p$, and a matrix $\mathbf{\Theta} \in \mathbb{R}^{n \times K}$ such that

$$m_{ij} = f_j\left(\boldsymbol{\theta}_{i,\cdot}\right), i = 1, \ldots, n, j = 1, \ldots, p, \tag{4.3}$$

where m_{ij} is the (i, j) entry of M and $\mathbf{\Theta} \in \mathbb{R}^{n \times K}$ is a matrix (with $\boldsymbol{\theta}_{i,\cdot}$ indicating the ith row vector). Here, $\mathbf{\Theta}$ gives an embedding of our observations from its original p-dimensional space into a K-dimensional space ($K \leq p$). The set of functions $\{f_j\}_{j=1}^p \subset \mathcal{F}$ identifies how to map our embedding in $\mathbb{R}^K$ back to $\mathbb{R}^p$.

In classical matrix completion setting, where we assume M is low-rank, nuclear norm penalized empirical risk minimization is often used to estimate M [3, 22, 78]; more specifically, the estimator is obtained by,

$$\arg\min_M \left\{ N^{-1} \sum_{t=1}^{N} (y_t - \langle X_t, M \rangle)^2 + \lambda \|M\|_* \right\}, \tag{4.4}$$

where λ is a regularization parameter which is used to balance the trade-off between fitting the unknown matrix using least squares and minimizing the nuclear norm $\|M\|_*$. This "matrix lasso" is known to have strong theoretical properties when M is low rank [3, 22, 78, 18]. However, in our scenario, M likely does not have low rank and previous work does not fully explain the effectiveness of the estimate from (4.4) in this setting.

While the estimator in (4.4) is simple and quite well known, it fails to exploit knowledge of the sampling scheme (which is often known or at least assumed to be known). To use the assumption that the mask matrices $\{X_t\}_{t=1}^N$ are i.i.d. uniformly sampled from $\mathcal{X}$, we study a

slight modification to (4.4) described in [61]:

$$\widehat{M} \leftarrow \arg\min_{M} \left\{ \frac{1}{np}\|M\|_F^2 - \left\langle \frac{2}{N}\sum_{t=1}^{N} y_t X_t, M \right\rangle + \lambda\|M\|_* \right\} \tag{4.5}$$

After some simple manipulation, (4.5) can be further reduced to minimizing

$$\frac{1}{np}\|M - R\|_F^2 + \lambda\|M\|_*.$$

where $R = \frac{np}{N}\sum_{t=1}^{N} y_t X_t = \frac{np}{N} Y$. Thus, $\widehat{M}$, the solution to (4.5), is merely a singular-value soft-thresholding estimator:

$$\widehat{M} = \sum_{j=1}^{\mathrm{rank}(R)} (\Lambda_j(R) - \lambda np/2)_+ u_j(R) v_j(R)^T, \tag{4.6}$$

where $\Lambda_j(R)$ are the singular values and $u_j(R)$, $v_j(R)$ are the left and right singular vectors of R such that $R = \sum_{j=1}^{\mathrm{rank}(R)} \Lambda_j(R) u_j(R) v_j(R)^T$. [61] established the rate optimality of this estimator with respect to Frobenius-norm loss when M is low rank. In this paper, we aim to ultimately claim that $\widehat{M}$ in (4.5) is still a consistent and rate optimal estimator of M in the case that M is non-linearly embeddable, as long as K is small and the function class $\mathcal{F}$ is sufficiently smooth.

4.2.3 Approximation of Embeddable Matrices

Our goal is to show that the estimator obtained by (4.5) is consistent for the true underlying matrix M with respect to Frobenius-norm loss (and characterize the convergence rate), when M is non-linearly embeddable. To this end, we first show that M can be well approximated by a series of matrices with low (and only slowly growing) rank as long as the function class $\mathcal{F}$ is sufficiently smooth. More specifically, we will need the following condition for the function class $\mathcal{F}$.

Condition 4.1. *Given a function class $\mathcal{F}$. Let C_0 denote a fixed positive number. Suppose*

that for any $\epsilon > 0$, there exists a finite set of functions $\mathcal{F}_\epsilon = \{\psi_1, \psi_2, \ldots, \psi_{J(\epsilon)}\} \subset \mathcal{F}$, such that

$$\|\psi\|_\infty \leq C_0, \quad \text{for all } \psi \in \mathcal{F}_\epsilon, \tag{4.7}$$

and

$$\max_{f \in \mathcal{F}} \min_{\|\beta\|_2^2 \leq C_0} \left\| f - \sum_{l=1}^{J(\epsilon)} \beta_l \psi_l \right\|_\infty \leq \epsilon. \tag{4.8}$$

For each ϵ, we denote by $\mathcal{F}_\epsilon^*$ a set of minimal cardinality such that (4.7) and (4.8) hold. We let $J^*(\epsilon)$ denote the cardinality of $\mathcal{F}_\epsilon^*$.

For a function class $\mathcal{F}$, Condition 4.1 characterizes the minimal number of basis functions needed to uniformly approximate functions in $\mathcal{F}$ up to precision ϵ. In Section 4.3, we shall apply this condition to K-dimensional, L-th order differentiable functions, and show how this number scales as a function of ϵ.

Based on the above condition, we can establishes the existence of an approximation matrix which is sufficiently close to the true matrix M and has a bounded nuclear norm.

Lemma 4.3. *Suppose matrix $M \in \mathbb{R}^{n \times p}$ is $\mathcal{F}$-embeddable, and $\mathcal{F}$ satisfies Condition 4.1. Then, for any $\epsilon > 0$, there exists a matrix M^ϵ satisfying $\operatorname{rank}(M^\epsilon) = J^*(\epsilon) \leq \min(n, p)$ such that*

$$\|M^\epsilon - M\|_\infty \leq \epsilon. \tag{4.9}$$

Furthermore, the nuclear norm of M^ϵ is bounded: There exists $C_1 > 0$ (independent of ϵ) such that

$$\frac{1}{\sqrt{np}} \|M^\epsilon\|_* \leq C_1 J^*(\epsilon). \tag{4.10}$$

The proof is given in Appendix C.1. Note, for the $\mathcal{F}$ we consider later (restricted to smooth functions) we will show that $J^*(\epsilon) << \min(n, p)$. This parallels results in classical non-parametric regression where many function-spaces considered can be approximated uniformly with small error by linear combinations of relatively few basis functions [107].

4.3 Consistency

Using Lemma 4.1, it is relatively straightforward to evaluate the performance of our estimator $\widehat{M}$ in (4.5). The performance metric simplest to theoretically analyze is $N^{-1}\sum_{i=1}^{N}\left\langle X_i, \widehat{M}-M\right\rangle^2$. However, this criterion only evaluates the prediction error on the *observed* entries. This is unsatisfying as our ultimate goal is to recover the entire matrix. Thus, we instead aim to evaluate the performance of $\widehat{M}$ based on the metric $\frac{1}{np}\|\widehat{M}-M\|_F^2$. The following result gives an upper bound for the performance of our estimator $\widehat{M}$ in this metric.

Theorem 4.8. *Suppose we observe N pairs $\{(y_t, X_t)\}_{t=1}^N$ satisfying data generating model (4.2) where X_t are i.i.d. uniformly sampled from $\mathcal{X}$. Assume the true matrix $M \in \mathbb{R}^{n\times p}$ is $\mathcal{F}$-embeddable where $\mathcal{F}$ satisfies Condition 4.1. Further suppose that $N \geq (n\wedge p)\log^2(n+p)$. Then there exists a constant $C_2 > 0$ (that only depends on σ and $\|M\|_\infty$) such that if we define the regularization parameter λ by*

$$\lambda = C_2\sqrt{\frac{\log(n+p)}{N(n\wedge p)}},$$

then, with probability at least $1-2(n+p)^{-1}$, the completion error of $\widehat{M}$ in (4.6) is bounded by

$$\frac{1}{np}\left\|\widehat{M}-M\right\|_F^2 \leq C_2^2\left(\frac{1+\sqrt{2}}{2}\right)^2\frac{(n\vee p)\log(n+p)}{N}J^*(\epsilon)+\epsilon^2, \tag{4.11}$$

for any $\epsilon > 0$. Here, $J^*(\epsilon)$ is the rank of the approximation matrix M^ϵ with $\|M-M^\epsilon\|_\infty \leq \epsilon$, which corresponds to the minimal cardinality of $\mathcal{F}^* \subset \mathcal{F}$ satisfying Condition 4.1.

The upper bound in Theorem 4.8 can be established by extending the results from [61]. The details of the proof are given in the Appendix C.2. The two terms on the right-hand-side of (4.11) clarify the trade-off between the approximation error, ϵ, and the cardinality of the minimal linear approximation set $\mathcal{F}^*$, $J^*(\epsilon)$. Our upper bound is consistent with the results in [61], where the error is decomposed into a misspecification error (ϵ^2) and a prediction error. Usually, when there is no misspecification, i.e., the true matrix M is low rank, the prediction

error is linearly related to the rank of M [23, 60, 18]. In our scenario, where the low-rank assumption is violated, the prediction error in (4.11) is linearly related to the rank of the approximation matrix.

Ideas similar to this occur in more traditional non-parametric estimation problems. For example, when using projection estimators in Hölder and Sobolev spaces, one of the main rate-optimal estimation approaches requires a truncated basis to be selected for projection that will grow with the sample size N [106]. However, in those examples, the number of basis vectors is a tuning parameter in the algorithm, and the set of basis functions must be selected in advance. Here, both the set of basis functions and the truncation level are rather just theoretical tools for analyzing the algorithm performance. In employing matrix completion, the analyst only needs to select λ.

We note that $N \geq (n \wedge p)\log^2(n+p)$ in the above Theorem 4.8 is a quite weak condition on the number of observations: N could satisfy this and still be far less than np. For the results of the latent space model in [24], they require at least $O\left(n^{\frac{2(K+1)}{K+2}}\right)$ entries to be observed out of n^2 entries to guarantee the consistency for recovering an $n \times n$ matrix. This implies that one needs to observe $O\left(n^{\frac{K}{K+2}}\right)$ entries out of n in each row, as compared to our much weaker requirement of $O\left(\log^2(n)\right)$ per row.

We now specialize our results to matrices that are $\mathcal{F}$-embeddable for $\mathcal{F}$ containing functions with bounded derivatives. This is a natural class of functions to work with (though one could alternatively work in a multivariate Sobolev or Hölder space).

Condition 4.2. *M is $\mathcal{F}$-embeddable, where $\mathcal{F}$ contains functions with uniformly bounded L-th order mixed partials (for some fixed $L > 0$). More formally, define $\mathcal{F}(L, \gamma, K)$, for $L, K \geq 1$ as the set of L-th order differentiable functions from $\mathbb{R}^K_{[0,1]}$ to $\mathbb{R}$ satisfying*

$$\left| \frac{\partial^L}{\partial x_1^{L_1} \cdots x_K^{L_K}} f(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x^0}} \right| \leq \gamma, \tag{4.12}$$

for all $\mathbf{x}^0 = (x_1^0, \ldots, x_K^0) \in \mathbb{R}^K_{[0,1]} \subset \mathbb{R}^K$ and all integers $L_1, \ldots, L_K$ satisfying $L_1 + \cdots + L_K = L$.

Now, additionally define the set

$$\begin{aligned}\mathcal{M}(L,\gamma,K) = \{M \in \mathbb{R}^{n\times p} \mid m_{ij} = f_j(\boldsymbol{\theta}_{i,\cdot}),\\ \text{with } f_j \in \mathcal{F}(L,\gamma,K),\ j \le p,\ \text{and } \boldsymbol{\theta}_{i,\cdot} \in \mathbb{R}^K_{[0,1]},\ i \le n\}\end{aligned} \tag{4.13}$$

This is the set of $F(L,\gamma,K)$ embeddable matrices, where the embedding lives in a compact space (for convenience we use the ℓ_∞ ball). Our formal condition here is that $M \in \mathcal{M}(L,\gamma,K)$.

Remark. In the above condition, we will often suppress the dependence on γ, and write $\mathcal{M}(L,K)$ and $\mathcal{F}(L,K)$: This is because γ does not affect the convergence rate of our estimator. Additionally, here we specify the domain of the embeddings to be $[0,1]^K$ for ease of exposition. This is actually general as we could rescale any compactly supported embedding to live in this interval.

Condition 4.2 imposes an additional constraint on our embedding: The underlying manifold on which our matrix lives should be smooth. Here smoothness is characterized by a number of bounded derivatives. As we will see, this function class engages well with Condition 4.1 in the sense that we are able to characterize $J^*(\epsilon)$ for the function class $\mathcal{F}(L,K)$. This is essentially a multivariate Hölder class, which has been widely used in the area of non-parametric estimation [106]. One could alternatively look at this as a multivariate Sobolev class under the sup-norm, $W^{L,\infty}(\mathbb{R}^K)$.

The following lemma gives the number of basis elements that is needed to linearly approximate a matrix satisfying the above condition, with bounded approximation error ϵ.

Lemma 4.4. *For the function class $\mathcal{F}(L,K)$ described in Condition 4.2, we have that Condition 4.1 is satisfied with $J^*(\epsilon) = O\left(\epsilon^{-K/L}\right)$.*

The proof of this lemma is given in Appendix C.3. Now, we can establish the final convergence result for smoothly embeddable matrices.

Theorem 4.9. *Under the same scenario and assumptions as in Theorem 4.8, assume further the $\mathcal{F}(K,L)$-embeddable matrix M satisfies Condition 4.2 for a given L and K. Then, the*

upper bound (4.11) *is optimized at* $\epsilon = \left(\frac{(n\vee p)\log(n+p)}{N}\right)^{\frac{L}{2L+K}}$, *resulting in*

$$\frac{1}{np}\left\|\widehat{M} - M\right\|_F^2 = O_P\left(\left[\frac{(n \vee p)\log(n+p)}{N}\right]^{\frac{2L}{2L+K}}\right). \tag{4.14}$$

The proof is given in Appendix C.4. This upper bound of the convergence rate of the MSE of $\widehat{M}$ is only based on the dimensions n and p of matrix M, the total number of observations N, as well as the degree of smoothness L and dimension of the embedding K. Previous work that assumed M was low-rank generally gave a rate of the form N^{-1} (n ∨ p)rank(M) log(n + p) [6, 60, 110]. In contrast, our upper bound does not rely on the rank of M. Instead, the role of rank(M) is replaced by L, and K. This result reaffirms that the standard matrix completion estimator based on nuclear norm minimization is consistent for matrices with low-dimensional non-linear structure. Perhaps more importantly, it also shows how the convergence rate depends on the degree of smoothness, and dimension of the manifold. This can be seen in the exponent on the RHS of (4.14): $\frac{2L}{2L+K}$. Increasing the degree of smoothness moves this exponent towards 1; increasing the dimension moves the exponent towards 0. This is analogous to more standard non-parametric regression problems in smooth hypothesis spaces where the minimax convergence rate for MSE looks analagous [106].

4.4 Minimax Lower Bound

In this section, we use information-theoretical methods to establish a lower bound on the estimation error for completing *non-linearly embeddable* matrices with *USR* entries when the latent embedding $\mathbf{\Theta}$ is K-dimensional and satisfies Condition 4.2. The rate we find in the lower bound matches the rate obtained by nuclear norm penalization in Theorem 4.9 up to a log-term. Thus our upper bound is sharp (up to a logarithmic factor), and, the nuclear-norm penalization based estimator given in (4.5) is rate-optimal (up to polylog) for this problem.

To derive the lower bound, we consider the underlying matrices $M \in \mathcal{M}(L, \gamma, K)$ as defined in (4.13), i.e., matrices that live in L-th order smooth, K dimensional manifolds.

Let $\mathbb{P}_M$ denote the probability distribution of the observations $\{(y_t, X_t)\}_{t=1}^N$ generated by model (4.2) with $\mathrm{E}(y_t|X_t) = \langle X_t, M\rangle$. We give a minimax lower bound of the $\|\cdot\|_F^2$-risk for estimating M in the following result.

Theorem 4.10. *For any given $L \geq 1$, $\gamma > 0$ and $K \geq 1$, let $\kappa := n/p$. Then, for some constant $A > 0$ that depends on K, L, γ, σ^2 and κ, the minimax risk for estimating M satisfies*

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{M}(L,\gamma,K)} \mathbb{P}_M \left(\frac{1}{np} \left\| \widehat{M} - M \right\|_F^2 > A \left(\frac{n \vee p}{N} \right)^{\frac{2L}{2L+K}} \right) \geq 1/2, \tag{4.15}$$

when $c_0^{-\frac{2L+K}{K}}(n \vee p) \leq N \leq c_0^{-\frac{2L+K}{K}} 0.48^{2L+K}(n \vee p) n^{\frac{2L+K}{K}}$ for some constant c_0 which depends on K, L, γ, σ^2 and κ.

The proof is given in the Appendix C.5. Comparing Theorem 4.10 to Theorem 4.9, we see that the lower bound matches the upper bound (4.14) up to a logarithmic factor. This shows that the estimator given by (4.5) is actually an optimal estimator (up to a log term) for this non-linear low-dimensional matrix completion regime.

We note that the requirement $N = O\left((n \vee p) n^{\frac{2L+K}{K}}\right)$ in Theorem 4.10 is a bit unusal. It comes from a technical constraint in our proof, required to construct a suitably large packing set. This may just be an artifact of our proof technique, and not innate to the problem. Recall that the upper bound holds as long as $N \geq (n \vee p)\log^2(n+p)$, so there is a large regime where the assumption required for our upper and lower bounds overlap.

4.5 Simulation Study

In this section, we empirically evaluate the effectiveness of matrix completion using the soft-thresholding estimator $\widehat{M}$ in (4.6) for noisy incomplete matrices which are generated from low-dimensional non-linear embeddings. (These matrices are full rank, even though they are generated from low-dimensional non-linear embeddings). Here, we only show the case of univariate embedding ($K = 1$) and aim to empirically evaluate how the Frobenius error $\frac{1}{np}\left\|\widehat{M} - M\right\|_F^2$ changes with the dimension (n) when $n = p$. We examine scenarios where the

non-linear embeddings are of different orders of smoothness.

The underlying matrices are generated as described in (4.3): $m_{ij} = f_j(\boldsymbol{\theta}_{i,\cdot})$ for $i = 1, \ldots, n$ and $j = i, \ldots, p$. In particular, to make sure that Conditions 4.1 and 4.2 are satisfied, we generate f_j as

$$f_j(x) = \sum_{b=1}^{\infty} \beta_b \psi_b(x),$$

where $\psi_b(x)$ are orthonormal bases in $L_2[0,1]$ defined by:

$$\begin{aligned} \psi_1(x) &= 1, \\ \psi_{2b}(x) &= \sqrt{2}\cos(2\pi b x), \\ \psi_{2b+1}(x) &= \sqrt{2}\sin(2\pi b x). \end{aligned}$$

Meanwhile, to set up the order of smoothness L and make sure that $\beta_b \psi_b(x)$ vanishes with b, we sample the coefficients β_b from a uniform distribution:

$$\beta_b \sim_{i.i.d} U\left[-b^{-(L+1)}, b^{-(L+1)}\right], \quad b = 1, 2, \ldots.$$

In this way, we can guarantee that $\sum_{b=1}^{\infty} b^{2L}\beta_j^2 < \infty$. Thus, f_j is a function whose Lth order derivative is $O_p(1)$.

In this simulation, for computational reasons, we actually use only the first 100 basis vectors $f_j(x) = \sum_{b=1}^{100} \beta_b \psi_b(x)$. The underlying embeddings $\boldsymbol{\theta}_{i,\cdot} \in \mathbb{R}$ are also i.i.d. sampled from a uniform distribution $U(0,1)$ for $i = 1, \ldots, n$. We set the missingness rate to $\nu = 0.3$: The total number of observed entries is $N = (1-\nu)np$. The observed entries are $y_t = \langle X_t, M \rangle + \xi_t$, where X_t are uniformly sampled from $\mathcal{X}$ and the error terms are independently Gaussian distributed $\xi_t \sim_{i.i.d.} N(0,1)$. We generate random data sets $\{(y_t, X_t)\}_{t=1}^N$ of size $n \in \{500, 1000, 2000, 3000, 5000\}$ and estimate M. We run 100 simulations for each size. To select λ, instead of using cross-validation, here we consider an oracle procedure: For each simulation, we estimate the MSE for a set of λ values and select the λ that minimizes the MSE. We report this MSE of the estimated matrix $\widehat{M}$ and the corresponding λ.

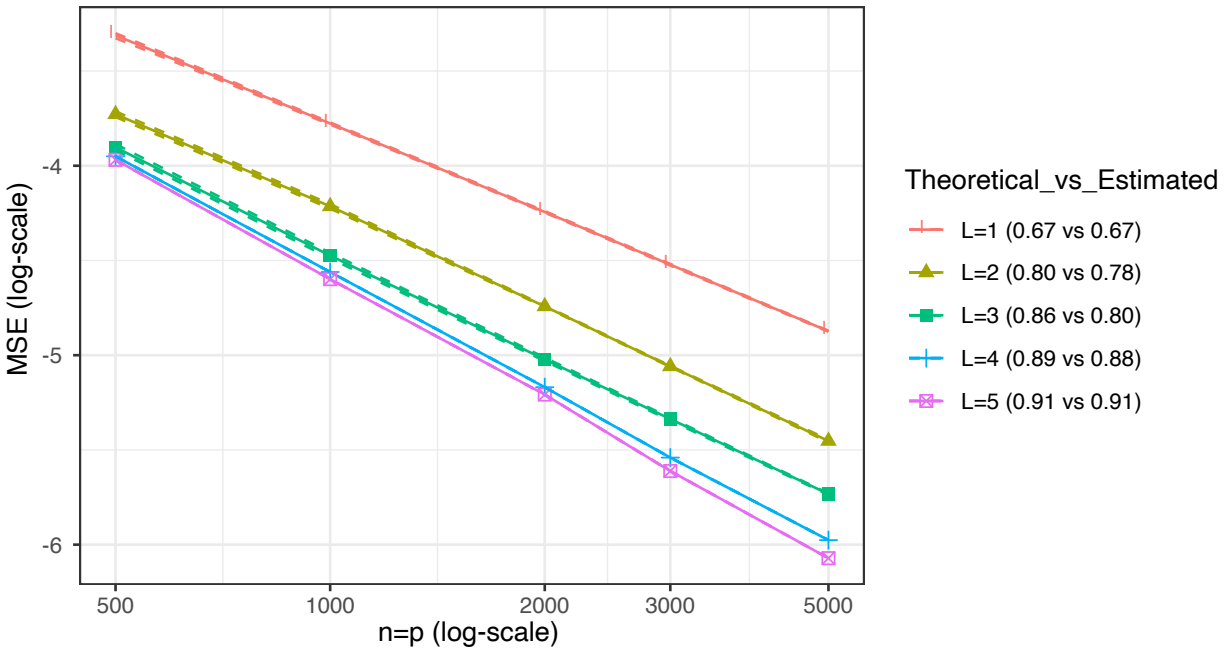

Figure 4.1: Theoretical rate vs. empirical rate (in log scale) of the mean squared errors as a function of sample size. The underlying matrices M are generated by f with different orders (L) of smoothness. The low-rank embedding is one-dimensional ($K = 1$). We regress log(MSE) on log(n), and compare the theoretical slopes (left) with the empirical slopes (right). For each smoothness level, L, we also obtain the 95% confidence regions using bootstrap (dash lines).

Figure 4.1 shows the results of estimating M generated by non-linear embeddings with different orders of smoothness, L. Since $N = (1-\nu)np$, the convergence rate in (4.14) reduces to $O_P\left(\left[\frac{\log(2n)}{n}\right]^{\frac{2L}{2L+K}}\right)$. The log term inside is negligible as n increases. Hence, if we regress log(MSE) on log(n), the absolute value of slope should be roughly about $\frac{2L}{2L+1}$ ($K = 1$ in this simulation). We increase the order of smoothness of f from $L = 1$ to $L = 5$. For these values of L, the expected absolute value of the slope should be 0.67, 0.80, 0.86, 0.89, and 0.91. The rates from our simulations are respectively 0.67, 0.78, 0.80, 0.88, and 0.91. There is generally strong agreement between theoretical and empirical results except for the setting of $L = 3$. We hypothesize that this is due to finite sample issues.

4.6 Discussion

Nuclear-norm based matrix completion methods were originally developed for scenarios where the underlying mean matrix has low rank. In this chapter, we present theoretical results to explain the effectiveness of matrix completion in applications where the underlying mean matrix is not low rank, but instead lives in a low-dimensional smooth manifold.

Our results show that, in such scenarios, nuclear-norm regularization can still result in a procedure that is minimax rate optimal (up to a log factor) for recovering the underlying mean matrix. In particular, we give upper bounds on the rate of convergence as a function of the number of rows, columns, and observed entries in the matrix, as well as the smoothness, and dimension of the embeddings. We additionally give matching minimax lower bounds (up to a logarithmic factor) for this problem. These bounds appear analogous to the minimax rate in the case of standard non-parametric regression.

Our theoretical results relate the error bounds to the smoothness and dimension of the non-linear embedding; however, the technical proof does not provide a way to figure out the explicit form of the hidden embeddings, which may be interesting in practice, e.g., for dimension reduction. Modifying the original matrix completion method in order to estimate the hidden embeddings may be an important direction of future research.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Proof of Theorem 2.1

A simple version of proof of theorem 1 is to directly calculate the discrepancy between $\widehat{\Psi}_1$ and $\Psi_1(P)$, which is

$$\begin{aligned}
&\mathbb{P}_n \left[y - \widehat{\mu}_Y(x)\right] \left[z - \widehat{\mu}_Z(x)\right] - \Psi_1(P) \\
&= \mathbb{P}_n \left\{ \left[y - \widehat{\mu}_Y(x)\right] \left[z - \widehat{\mu}_Z(x)\right] - \widehat{\Psi}_1 \right\} + \widehat{\Psi}_1 - \Psi_1(P) \\
&= [\mathbb{P}_n - P]\widehat{D}^{(1)} + P\widehat{D}^{(1)} + \widehat{\Psi}_1 - \Psi_1(P) \\
&= [\mathbb{P}_n - P]D_P^{(1)} + [\mathbb{P}_n - P][\widehat{D}^{(1)} - D_P^{(1)}] + P\left[(\widehat{\mu}_Y - \mu_{P,Y})(\widehat{\mu}_Z - \mu_{P,Z})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}[y_i - \mu_{P,Y}(x_i)][z_i - \mu_{P,Z}(x_i)] + o_P(n^{-1/2})
\end{aligned} \tag{A.1}$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ and $Pf = \int f dP$. The last equation in (A.1) holds when Assumption 1 holds. In this case, we have the asymptotic linearity and normality. As an alternative route to prove the theorem we can apply standard semi-parametric tools: We calculate the efficient influence function and then consider a first order asymptotic expansion to show that the theoretically optimal plug-in estimator of $\Psi_1(P)$ has exactly the same form as, so-called one-step estimator. Thus, it will naturally enjoy the good properties including asymptotic consistency and normality.

For $\Psi_2(P)$ however, our simple direct approach will not work, so instead we need to apply those semi-parametric tools.

A.2 Proof of Theorem 2.1 by Semi-parametric Theory

For a distribution $P \in \mathcal{M}$, let p denote the density with respect to a dominant measure ν. We define a parametric sub-model $p_\theta(u) := [1+\theta h(u)]p(u)$ with $h(u) \in L_2(P)$, where $\mathrm{E}\{h(u)\} = 0$, $\sup_u |h(u)| < \infty$, and θ is sufficiently small, such that $p_\theta \geq 0$ and $\int p_\theta(u)d\nu(u) = 1$. Upon inspection we see that this parametric sub-model is centered at P with score $s_\theta(u)|_{\theta=0} = \frac{\partial}{\partial\theta}\log p_\theta(u)|_{\theta=0} = h(u)$. In this framework, our statistical functional $\Psi_1(P)$ is called pathwise differentiable at P with efficient influence function $D_P^{(1)}$ [14], if

$$\frac{\partial}{\partial\theta}\Psi_1(P_\theta)(u)\bigg|_{\theta=0} = \int D_P^{(1)}(u)h(u)dP(u) \tag{A.2}$$

.

Consider observed data consisting of an independent and identically distributed sample of $o = (y, z, x) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ drawn from distribution P. Then, the corresponding parametric submodel p_θ can be conditionally decomposed into

$$p_\theta(o) = p_{\theta,h_y}(y|z,x)p_{\theta,h_z}(z|x)p_{\theta,h_x}(x), \tag{A.3}$$

with score at the origin

$$\begin{aligned} s_\theta(o)|_{\theta=0} &= s_\theta(y|z,x)|_{\theta=0} + s_\theta(z|x)|_{\theta=0} + s_\theta(x)|_{\theta=0} \\ &= h_y(y;z,x) + h_z(z;x) + h_x(x). \end{aligned} \tag{A.4}$$

For simplicity, we use h_y, h_z, and h_x to represent $h_y(y;z,x)$, $h_z(z;x)$, and $h_x(x)$ respectively. We first show how to obtain the efficient influence function $D_P^{(1)}$ stated in Theorem 1. Let $\psi_P(x)$, $\mu_{P,Y}(x)$, and $\sigma^2_{P,Y}(x)$ denote the conditional covariance $\mathrm{Cov}_P(Y,Z|X)$, conditional mean $\mathrm{E}_P(Y|X=x)$, and conditional variance $\mathrm{Var}_P(Y|X=x)$ evaluated under true model P, while $\psi_\theta(x)$, $\mu_{\theta,Y}(x)$, and $\sigma^2_{\theta,Y}(x)$ are evaluated under sub-model P_θ. The expected conditional

covariance $\Psi_1(P)$ in (2.2) evaluated on $P_\theta|_{\theta=0}$ is

$$\begin{aligned}\frac{\partial}{\partial\theta}\Psi_1(P_\theta)\bigg|_{\theta=0} &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}}\psi_\theta(x)dP_\theta(x)\bigg|_{\theta=0} \\ &= \int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right)dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}}\psi_\theta(x)\frac{\partial}{\partial\theta}\log p_\theta(x)dP_\theta(x)\bigg|_{\theta=0} \\ &= \int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right)dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}}\psi_1(x)h_x dP(x)\end{aligned} \tag{A.5}$$

Also, we have $\psi_\theta(x) = \mu_{\theta,YZ}(x) - \mu_{\theta,Y}(x)\mu_{\theta,Z}(x)$, where

$$\begin{aligned}\mu_{\theta,YZ}(x) &= \int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp_\theta(y|z,x)p_\theta(z|x)dydz \\ &= \int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y|z,x)(1+\theta h_y)p(z|x)(1+\theta h_z)dydz \\ &= \mu_{P,YZ}(x) + \theta\int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y,z|x)(h_y+h_z)dydz \\ &\quad + \theta^2\int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y,z|x)h_y h_z dydz \\ &= \mu_{P,YZ}(x) + \theta\,\mathrm{E}\left[YZ(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[YZh_y h_z|X=x\right].\end{aligned} \tag{A.6}$$

and similarly,

$$\mu_{\theta,Y}(x) = \mu_{P,Y}(x) + \theta\,\mathrm{E}\left[Y(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[Yh_y h_z|X=x\right],$$

$$\mu_{\theta,Z}(x) = \mu_{P,Z}(x) + \theta\,\mathrm{E}\left[Z(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[Zh_y h_z|X=x\right].$$

We then get that

$$\begin{aligned}\int_{\mathcal{X}} \frac{\partial}{\partial\theta}\psi_\theta(x)dP_\theta(x)\Big|_{\theta=0} &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\mu_{\theta,YZ}(x) - \mu_{\theta,Y}(x)\frac{\partial}{\partial\theta}\mu_{\theta,Z}(x) - \mu_{\theta,Z}(x)\frac{\partial}{\partial\theta}\mu_{\theta,Y}(x)dP_\theta(x)\Big|_{\theta=0} \\ &= \int_{\mathcal{X}} \mathrm{E}[YZ(h_y+h_z)|X=x] - \mu_{P,Z}(x)\,\mathrm{E}[Y(h_y+h_z)|X=x] \\ &\quad - \mu_{P,Y}(x)\,\mathrm{E}[Z(h_y+h_z)|X=x]dP(x) \\ &= \mathrm{E}\left\{[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X)) - \Psi_1(P)]\,(h_y+h_z+h_x)\right\}. \end{aligned} \tag{A.7}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial\theta}\Psi_1(P_\theta)\Big|_{\theta=0} &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right) dP_\theta(x)\Big|_{\theta=0} + \int_{\mathcal{X}} \Psi_1(x)h_x dP(x) \\ &= \mathrm{E}\left\{[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X)) - \Psi_1(X)]\,(h_y+h_z+h_x)\right\} + \mathrm{E}[\Psi_1(X)h_x] \\ &= \mathrm{E}\left\{[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X))]\,(h_y+h_z+h_x)\right\}. \end{aligned} \tag{A.8}$$

which gives us the efficient influence function $D_P^{(1)}(o) = (y-\mu_{P,Y}(x))(y-\mu_{P,Z}(x)) - \Psi_1(P)$ in Theorem 1. We note that, in above equation, we use the fact that $\int f(y,z)h_x dP(o) = 0$ and $\int g(x)(h_y+h_z)dP(o) = 0$ where $f(y,z)$ is an arbitrary function which does not depend on x and $g(x)$ is an arbitrary function which depends only on x.

Now, we can use the efficient influence function to obtain the so-called "one-step estimator". Consider an asymptotic von-mises expansion of Ψ_1 centered at the true P and evaluated at some $P^* \in \mathcal{M}$ [41]. Then we have that

$$\begin{aligned}\Psi_1(P^*) - \Psi_1(P) &= (P^*-P)D_{P^*}^{(1)} + R_1(P^*,P) \\ &= -PD_{P^*}^{(1)} + R_1(P^*,P), \end{aligned} \tag{A.9}$$

where $R(P^*,P)$ is a second order remainder term and we use the fact that $PD_{P^*}^{(1)} = 0$. We

can now plug in an estimated distribution $\widehat{P}_n$, and use a bit of algebra to show

$$\begin{aligned}\Psi_1(\widehat{P}_n) - \Psi_1(P) &= -\mathbb{P}_n\widehat{D}^{(1)} + (\mathbb{P}_n - P)\widehat{D}^{(1)} + R_1(\widehat{P}_n, P) \\ &= -\mathbb{P}_n\widehat{D}^{(1)} + (\mathbb{P}_n - P)D_P^{(1)} + (\mathbb{P}_n - P)\left(\widehat{D}^{(1)} - D_P^{(1)}\right) + R_1(\widehat{P}_n, P),\end{aligned} \tag{A.10}$$

The second term above is the linear term evaluated at the truth, with mean zero. Ideally, we can find an estimator for Ψ_1 such that this term can dominate the asymptotic performance of $\Psi_1(\widehat{P}_n)$. The third and fourth one are respectively an empirical process term and second-order remainder term, which can be shown to be negligible under certain conditions on $\widehat{P}_n$. That is, they both converge to 0 faster than the linear term as $n \to \infty$. However, we see that the term $\mathbb{P}_n\widehat{D}^{(1)}$ is the source of the irregular behavior of $\Psi_1(\widehat{P}_n)$ and can often cause non-ignorable bias. Hence, this expansion motivates us to find a proper way to cancel the effects of $\mathbb{P}_n\widehat{D}^{(1)}$ and give the proposed one-step estimator for $\Psi_1(P)$

$$\begin{aligned}\widehat{\Psi}_{1,onestep} &= \Psi_1(\widehat{P}_n) + \mathbb{P}_n\widehat{D}^{(1)} \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu}_Y(x_i))(y_i - \widehat{\mu}_Z(x_i)),\end{aligned} \tag{A.11}$$

which is coincidentally the same as the proposed theoretically optimal plug-in estimator $\widehat{\Psi}_1$ in our paper. Therefore, according to (A.10), we can also obtain the asymptotic linearity of $\widehat{\Psi}_1$:

$$\widehat{\Psi}_1 - \Psi_1(P) = \frac{1}{n}\sum_{i=1}^{n} D_P^{(1)}(o_i) + o_P(n^{-1/2}),$$

as long as the empirical process $(\mathbb{P}_n - P)\left(\widehat{D}^{(1)} - D_P^{(1)}\right)$ and the second-order remainder term $R_1(\widehat{P}_n, P)$ are negligible. By Assumption 1 we have that $(\mathbb{P}_n - P)\left(\widehat{D}^{(1)} - D_P^{(1)}\right) = o_P(n^{-1/2})$. Thus, we only need to prove $R_1(\widehat{P}_n, P) = o_P(n^{-1/2})$. For any $P^* \in \mathcal{M}$, the

remainder is

$$\begin{aligned} R_1(P^*, P) &= \Psi_1(P^*) - \Psi_1(P) + PD^{(1)}_{P^*} \\ &= P\left\{(Y - \mu_{P^*,Y}(X))(Z - \mu_{P^*,Z}(X))\right\} - P\left\{(Y - \mu_{P,Y}(X))(Z - \mu_{P,Z}(X))\right\} \\ &= P\left\{(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))\right\}. \end{aligned} \tag{A.12}$$

Hence, as long as $\mu_{P^*,Y}(X) - \mu_{P,Y}(X)$ and $\mu_{P^*,Z}(X) - \mu_{P,Z}(X)$ both converge to zero at $o_P(n^{-1/4})$, we have $R_1(P^*, P) = o_P(n^{-1/2})$. That is to say, under Assumptions 1, the asymptotic linearity of $\widehat{\Psi}_1$ holds. By the central limit theorem, we can further derive the asymptotic normality of $\widehat{\Psi}_1$, i.e.

$$\sqrt{n}[\widehat{\Psi}_1 - \Psi_1(P)] \to_d N[0, \sigma_1^2(P)], \tag{A.13}$$

where $\sigma_1^2(P) = \int [D_P^{(1)}(o)]^2 dP(o)$. This completes the proof of Theorem 1

A.3 Proof of Theorem 2.3

As in (A.9), we can also show that the naive plug-in estimator $\widehat{\Psi}_{2,naive}$ is asymptotically biased, which can be corrected by adding the irregular bias. Let $\phi_\theta(x) = \mathrm{Cor}(Y, Z|X)$ under P_θ and $\phi(x) = \phi_\theta|_{\theta=0}$. Then, $\phi_\theta(x) = \frac{\psi_\theta(x)}{g_\theta(x)}$, where $g_\theta(x) = \sqrt{\sigma^2_{\theta,Y}(x)\sigma^2_{\theta,Z}(x)}$. The conditional variance under P_θ can be expanded as follows,

$$\begin{aligned} \sigma^2_{\theta,Y}(x) &= \mu_{\theta,Y^2}(x) - \mu^2_{\theta,Y}(x) \\ &= \mu_{Y^2}(x) + \theta\,\mathrm{E}\left[Y^2(h_y + h_z)|X = x\right] + \theta^2\,\mathrm{E}\left[Y^2 h_y h_z|X = x\right] \\ &\quad - \left\{\mu_{P,Y}(x) + \theta\,\mathrm{E}\left[Y(h_y + h_z)|X = x\right] + \theta^2\,\mathrm{E}\left[Y h_y h_z|X = x\right]\right\}^2 \\ &= \sigma^2_{P,Y}(x) + \theta\,\mathrm{Cov}(Y, Y(h_y + h_z)|X = x) + \theta^2\,\mathrm{Cov}(Y, Y h_y h_z|X = x) \\ &\quad - \mu_{P,Y}(x)\left\{\theta\,\mathrm{E}[Y(h_y + h_z)|X = x] + \theta^2\,\mathrm{E}[Y h_y h_z|X = x]\right\} \\ &\quad - \left\{\theta\,\mathrm{E}[Y(h_y + h_z)|X = x] + \theta^2\,\mathrm{E}[Y h_y h_z|X = x]\right\}^2. \end{aligned} \tag{A.14}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\sigma^2_{\theta,Y}(x)\bigg|_{\theta=0} = \mathrm{Cov}(Y, Y(h_y+h_z)|X=x) - \mu_{P,Y}(x)\,\mathrm{E}[Y(h_y+h_z)|X=x] \tag{A.15}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\sigma^2_{\theta,Z}(x)\bigg|_{\theta=0} = \mathrm{Cov}(Z, Z(h_y+h_z)|X=x) - \mu_{P,Z}(x)\,\mathrm{E}[Z(h_y+h_z)|X=x]. \tag{A.16}$$

Thus, the derivative of $\Psi_2(P_\theta)$ with respect to θ at $\theta = 0$ is

$$\begin{aligned}\frac{\partial}{\partial\theta}\Psi_2(P_\theta)\bigg|_{\theta=0} &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\phi_\theta(x)\right) dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}} \phi(x)h_x dP(x) \\ &= \int_{\mathcal{X}} \frac{\psi'_\theta(x)g_\theta(x) - \psi_\theta(x)g'_\theta(x)}{g^2_\theta(x)} dP_\theta(x)\bigg|_{\theta=0} + \mathrm{E}[\phi(X)h_x],\end{aligned} \tag{A.17}$$

where

$$\begin{aligned}\psi'_\theta(x)|_{\theta=0} = \mathrm{E}[YZ(h_y+h_z)|X=x] &- \mu_{P,Z}(x)\,\mathrm{E}[Y(h_y+h_z)|X=x] \\ &- \mu_{P,Y}(x)\,\mathrm{E}[Z(h_y+h_z)|X=x],\end{aligned} \tag{A.18}$$

and

$$\begin{aligned}g'_\theta(x) = \frac{1}{2g(x)}&\left\{\mathrm{E}[Y^2(h_y+h_z)|X] - 2\mu_{P,Y}(x)\,\mathrm{E}[Y(h_y+h_z)|X]\right\}\sigma^2_{P,Z}(x) \\ &+ \frac{1}{2g(x)}\left\{\mathrm{E}[Z^2(h_y+h_z)|X] - 2\mu_{P,Z}(x)\,\mathrm{E}[Z(h_y+h_z)|X]\right\}\sigma^2_{P,Y}(x).\end{aligned} \tag{A.19}$$

Plugging in (A.18) and (A.19), (A.17) becomes

$$\begin{aligned}
&\int_{\mathcal{X}} \frac{\psi_\theta'(x)g_\theta(x) - \psi_\theta(x)g_\theta'(x)}{g_\theta^2(x)} dP_\theta(x)\bigg|_{\theta=0} + \mathrm{E}[\phi(X)h_x] \\
&= \mathrm{E}\left\{ \left[\frac{(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X))}{g(X)} - \phi(X) \right] (h_y + h_z + h_x) \right\} \\
&\quad - \mathrm{E}\left\{ \phi(X) \left[\frac{(Z-\mu_{P,Z}(X))^2}{2\sigma_{P,Z}^2(X)} + \frac{(Y-\mu_{P,Y}(X))^2}{2\sigma_{P,Y}^2(X)} - 1 \right] (h_y + h_z + h_x) \right\} + \mathrm{E}[\phi(X)h_x] \\
&= \mathrm{E}\left\{ \left[\frac{(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X))}{g(X)} - \phi(X) \left(\frac{(Z-\mu_{P,Z}(X))^2}{2\sigma_{P,Z}^2(X)} + \frac{(Y-\mu_{P,Y}(X))^2}{2\sigma_{P,Y}^2(X)} - 1 \right) - \Psi_2(P) \right] (h_y + \right.
\end{aligned} \tag{A.20}$$

Therefore, the efficient influence function of $\Psi_2(P)$ is

$$D_P^2(o) = \frac{(y-\mu_{P,Y}(x))(z-\mu_{P,Z}(x))}{g(x)} - \phi(x) \left(\frac{(z-\mu_{P,Z}(x))^2}{2\sigma_{P,Z}^2(x)} + \frac{(y-\mu_{P,Y}(x))^2}{2\sigma_{P,Y}^2(x)} - 1 \right) - \Psi_2(P). \tag{A.21}$$

Thus, the one-step estimator of $\Psi_2(P)$ according to (A.10) is

$$\begin{aligned}
\widehat{\Psi}_2 &= \Psi_2(\widehat{P}_n) + \mathbb{P}_n \widehat{D}^{(2)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(y_i - \widehat{\mu}_Y(x_i))(z_i - \widehat{\mu}_Z(x_i))}{\sqrt{\widehat{\sigma}_Y^2(x_i)\widehat{\sigma}_Z^2(x_i)}} \right. \\
&\quad \left. - \frac{\widehat{\mu}_{YZ}(x_i) - \widehat{\mu}_Y(x_i)\widehat{\mu}_Z(x_i)}{\sqrt{\widehat{\sigma}_Y^2(x_i)\widehat{\sigma}_Z^2(x_i)}} \left[\frac{(y_i - \widehat{\mu}_Y(x_i))^2}{2\widehat{\sigma}_Y^2(x_i)} + \frac{(z_i - \widehat{\mu}_Z(x_i))^2}{2\widehat{\sigma}_Z^2(x_i)} - 1 \right] \right\}.
\end{aligned} \tag{A.22}$$

The second-order remainder of $\Psi_2(P^*)$ is

$$
\begin{aligned}
R_2(P^*, P) &= \Psi_2(P^*) - \Psi_2(P) + PD^{(2)}_{P^*} \\
&= P\left\{\frac{(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))}{g_{P^*}(X)}\right\} \\
&\quad - P\left\{\frac{\mathrm{Corr}_{P^*}(Y,Z|X)}{2\sigma^2_{P^*,Y}(X)}\,(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))^2\right\} \\
&\quad - P\left\{\frac{\mathrm{Corr}_{P^*}(Y,Z|X)}{2\sigma^2_{P^*,Z}(X)}\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))^2\right\} \\
&\quad + P\left\{\frac{\mathrm{Cov}_{P^*}(Y,Z|X) - \mathrm{Cov}_{P}(Y,Z|X)}{g_{P^*}(X)}\left(\frac{\sigma^2_{P^*,Y}(X) - \sigma^2_{P,Y}(X)}{2\sigma^2_{P^*,Y}(X)} + \frac{\sigma^2_{P^*,Z}(X) - \sigma^2_{P,Z}(X)}{2\sigma^2_{P^*,Z}(X)}\right)\right\} \\
&\quad - P\left\{f_1(X)(\sigma_{P^*,Z}(X) - \sigma_{P,Z}(X))^2 - f_2(X)(\sigma_{P^*,Y}(X) - \sigma_{P,Y}(X))(\sigma_{P^*,Z}(X) - \sigma_{P,Z}(X))\right. \\
&\quad \left. + f_3(X)(\sigma_{P^*,Y}(X) - \sigma_{P,Y}(X))^2\right\},
\end{aligned}
\tag{A.23}
$$

where $\{f_i\}_{i=1}^3$ are some functions depending only on X. Hence, to make $R_2(\widehat{P}_n, P)$ converges to zero at $o_P(n^{-1/2})$, we have to guarantee that every item in (A.23) converges to zero at $o_P(n^{-1/2})$, which includes $\int \left(\mathrm{Cov}_{\widehat{P}_n}(Y,Z|x) - \mathrm{Cov}_P(Y,Z|x)\right)\left(\sigma^2_{P^*,Y}(x) - \sigma^2_{P,Y}(x)\right) dP(x)$, $\int \left(\widehat{\sigma}_Y(x) - \sigma_{P,Y}(x)\right)^2 dP(x)$, $\int \left(\widehat{\sigma}_Z(x) - \sigma_{P,Z}(x)\right)^2 dP(x)$. Then, we have the asymptotic linearity of $\Psi_2(P)$ by (A.10),

$$
\widehat{\Psi}_2 - \Psi_2(P) = \frac{1}{n}\sum_{i=1}^{n} D^{(2)}_P(o_i) + o_P(n^{-1/2}),
$$

and the asymptotic normality

$$
\sqrt{n}[\widehat{\Psi}_2 - \Psi_2(P)] \to_d N[0, \sigma_2^2(P)], \tag{A.24}
$$

where $\sigma_2^2(P) = \int [D^{(2)}_P(o)]^2 dP(o)$. This completes the proof of Theorem 3.

A.4 Proof of Theorem 2.2

Before proving Theorem 2, we first show that $-1 \leq \Psi_3(P) \leq 1$ for all P. By applying Cauchy-Schwartz and Jensen's inequality, we get that

$$\mathrm{Cov}^2(Y, Z|X) \leq \mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)$$

$$|\,\mathrm{E}[\mathrm{Cov}(Y, Z|X)]| \leq \mathrm{E}\sqrt{\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)} \leq \sqrt{\mathrm{E}[\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)]}$$

$$|\Psi_3(P)| = \left|\frac{\mathrm{E}[\mathrm{Cov}(Y, Z|X)]}{\sqrt{\mathrm{E}[\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)]}}\right| \leq 1,$$

which has the same range as the correlation.

The efficient influence function of $\Psi_3(P)$ can be easily derived from what we have developed for $\Psi_1(P)$ by delta method [98]. Recall that expected conditional covariance has efficient influence function $D_P^{(1)}(o)$, So the efficient influence function for the expected conditional variance $V_Y(P)$ is $D_P^{V_Y}(o) = (y - \mu_{P,Y}(x))^2$. Let $g(u, v, w) = \frac{u}{\sqrt{vw}}$. Then we know that $\Psi_3(P) = g(\Psi_1(P), V_Y(P), V_Z(P))$ has efficient influence function

$$\begin{aligned} D_P^{(3)}(o) &= \nabla g(\Psi_1(P), V_Y(P), V_Z(P)) \times (D_P^{(1)}(o), D_P^{V_Y}(o), D_P^{V_Z}(o))^T \\ &= \frac{(y - \mu_{P,Y}(x))(z - \mu_{P,Z}(x))}{\sqrt{V_Y(P)V_Z(P)}} - \Psi_3(P)\left[\frac{(y - \mu_{P,Y}(x))^2}{2V_Y(P)} + \frac{(z - \mu_{P,Z}(x))^2}{2V_Z(P)}\right]. \end{aligned} \tag{A.25}$$

Thus, the one-step estimator is exactly the same as our theoretically optimal plug-in estimator, becausel of $\mathbb{P}_n\widehat{D}^{(3)} = \Psi_3(\widehat{P}_n) - \Psi_3(\widehat{P}_n)\left[\frac{\frac{1}{n}\sum(y_i - \mu_Y(x_i))^2}{2V_Y(\widehat{P}_n)} + \frac{\frac{1}{n}\sum(z_i - \mu_Z(x_i))^2}{2V_Z(\widehat{P}_n)}\right] = 0$. Then

the second order remainder of $\Psi_3(P^*)$ is

$$\begin{aligned} R_3(P^*, P) &= \Psi_3(P^*) - \Psi_3(P) + PD^{(3)}_{P^*} \\ &= \frac{P[(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))}{\sqrt{V_Y(P^*)V_Z(P^*)}} \\ &\quad - \frac{\Psi_3(P^*)}{2}\left[\frac{P(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))^2}{V_Y(P^*)} + \frac{P(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))^2}{V_Z(P^*)}\right] \\ &\quad + G(P^*, P) \end{aligned} \tag{A.26}$$

where

$$\begin{aligned} G(P^*, P) = \frac{\Psi_1(P^*) - \Psi_1(P)}{2\sqrt{V_Y(P^*)V_Z(P^*)}}&\left[\frac{(V_Y(P^*) - V_Y(P))^2}{V_Y(P^*)} + \frac{(V_Z(P^*) - V_Z(P))^2}{V_Z(P^*)}\right] \\ - \frac{\Psi_1(P)}{\sqrt{V_Y(P^*)V_Z(P^*)}}&\left[\frac{[\sqrt{V_Y(P)V_Z(P)} - \sqrt{V_Y(P^*)V_Z(P^*)}]^2}{\sqrt{V_Y(P^*)V_Z(P^*)V_Y(P)V_Z(P)}}\right. \\ &+ \frac{[\sqrt{V_Z(P^*)} - \sqrt{V_Z(P)}]^2}{2V_Z(P^*)} + \frac{[\sqrt{V_Y(P^*)} - \sqrt{V_Y(P)}]^2}{2V_Y(P^*)} \\ &\left. - \frac{[\sqrt{V_Y(P^*)} - \sqrt{V_Y(P)}]\sqrt{V_Z(P^*)} - \sqrt{V_Z(P)}]}{\sqrt{V_Y(P^*)V_Z(P^*)}}\right] \end{aligned}$$

Under Assumption 1, we have known that $\Psi_1(P^*) - \Psi_1(P) = o_P(n^{-1/2})$. Thus, $G(P^*, P) = o_P(n^{-1})$ and thus, $R_3(P^*, P) = o_P(n^{-1/2})$ is negligible. We can then obtain the asymptotical linearity and nonparametric efficiency of $\widehat{\Psi}_3$ as in Theorem 2.

A.5 Additional experiments for asymptotic performance

For $\Psi_1(P)$, $\Psi_2(P)$, $\Psi_3(P)$, we compare the efficient estimators proposed in paper, with their corresponding naive estimators (which should theoretically not be rate optimal):

$$\widehat{\Psi}_{1,naive} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{\mu}_{YZ}(x) - \widehat{\mu}_Y(x)\widehat{\mu}_Z(x)] \tag{A.27}$$

$$\widehat{\Psi}_{2,naive} = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{\mu}_{YZ}(x_i) - \widehat{\mu}_Y(x_i)\widehat{\mu}_Z(x_i)}{\sqrt{\left(\widehat{\mu}_{Y^2}(x_i) - \widehat{\mu}_Y^2(x_i)\right)\left(\widehat{\mu}_{Z^2}(x_i) - \widehat{\mu}_Z^2(x_i)\right)}} \tag{A.28}$$

$$\widehat{\Psi}_{3,naive} \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_{YZ}(x) - \widehat{\mu}_Y(x)\widehat{\mu}_Z(x)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{Y^2}(x_i) - \widehat{\mu}_Y^2(x_i)\right) \times \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{Z^2}(x_i) - \widehat{\mu}_Z^2(x_i)\right)}}. \tag{A.29}$$

A.5.1 Low-dimensional cases

We modify the setting of low-dimensional example in our paper slightly, by changing the underlying covariance structure of errors of Y and Z. In this case, we let

$$\vec{e}|X = (e_y, e_z)^T|X \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 + \frac{x}{4} \\ -0.5 + \frac{x}{4} & 1 \end{pmatrix}\right]. \tag{A.30}$$

The true value of $\Psi_1(P)$ is 0.25. The results are shown in Figure A.1: the naive estimator again does not have a bias converging to zero at $o_P(n^{-1/2})$ and we cannot obtain a valid confidence interval by bootstrapping. In fact, we can notice that bootstrap-based methods indeed fails quite spectacularly.

We also use the same pattern in low-dimensional setting described in our paper to evaluate the theoretically optimal plug-in & naive estimator of $\Psi_3(P)$. Figure A.2 shows the results. The empirical $\sqrt{n}$-scaled bias of our theoretically optimal estimator $\widehat{\Psi}_3$ goes toward zero which this is not the case for the naive estimator. The empirical variance of both methods stabilizes when scaled by n and the confidence interval of our optimal plug-in estimators converges to the nominal 95% as sample size increases. As expected, due to excess bias, the bootstrap interval based on the "naive" estimators performs poorly (with coverage actually converging to 0)

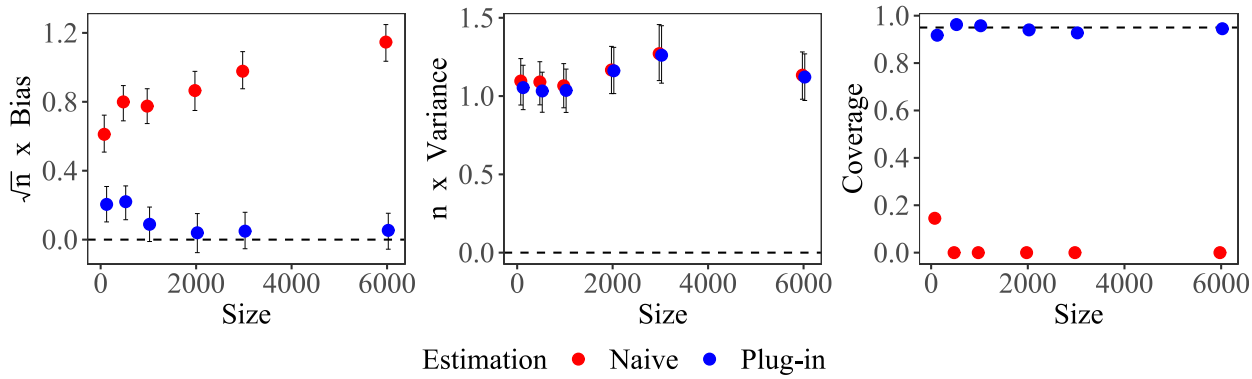

Figure A.1: Low dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_1(P)$. Conditional mean is estimated by local polynomial regression.

A.5.2 Moderate dimensional cases

In this setting, we generate the data from following mechanism.

$$Y = f_1(x_1, ..., x_8) + e_y, \quad Z = f_2(x_1, ..., x_8) + e_z, \tag{A.31}$$

where $X \sim N(0, I_8)$ and $\vec{e} = (e_y, e_z)^T \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right]$. Here the true value of expected conditional covariance is also $\Psi_1(P) = -0.5$. For each sample size $n \in \{300, 500, 2000, 4000, 6000, 8000, 10000\}$, we generated 400 datasets. Gradient boosting were used to estimate the conditional means $\mu_Y(x)$ and $\mu_Z(x)$ where hyper-parameters (number of trees, minimal node size and fraction of observations to sample) are tuned by a 5-fold cross validation. Since bootstrap-based approach fails to build the confidence interval and is computationally expensive. Here, we just include the Wald-type confidence interval of the optimal plug-in estimator. The results look similar to low-dimensional cases, see Figure A.3. As n increases, $\sqrt{n}$-scaled bias of the optimal plug-in estimator tends to zero while that of the naive estimator diverges. The variances go to a positive constant and the empirical

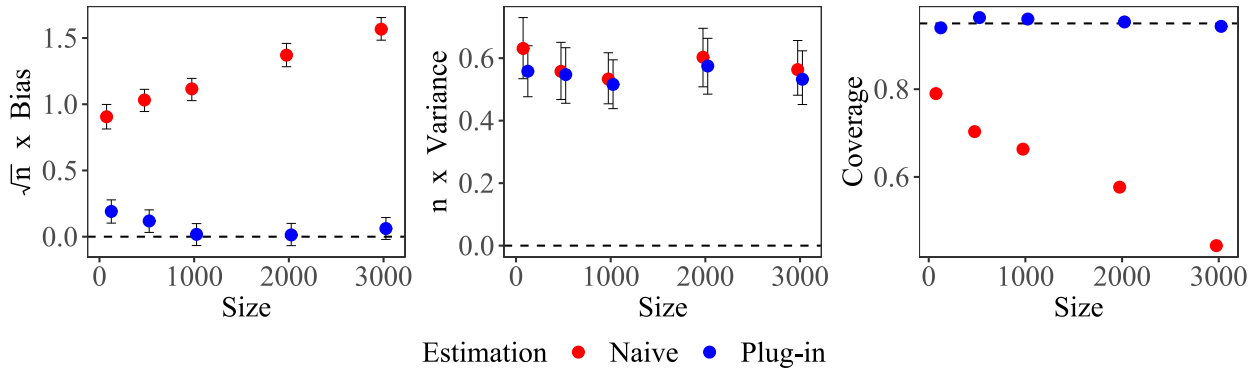

Figure A.2: Low dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_3(P)$. Conditional mean is estimated by local polynomial regression.

coverage obtained from asymptotic normality also works. In this setting, we may notice that bias of the optimal plug-in estimator converges to 0 more slowly but still gives reasonable interval estimates.

A.5.3 High-dimensional cases

We use the same setting with high-dimensional features to evaluate the performance of the scaled expected conditional covariance Ψ_3. The true parameter is $\Psi_3(P) = -0.5$. We generate random datasets of size $n \in \{500, 1000, 2000, 3000, 4000\}$ and estimate Ψ_3. The Lasso was used to estimating the conditional means $\mu_Y(x)$ and $\mu_Z(x)$ where the regularization parameter was tuned by a 5-fold cross validation. Again, the results are in-line with our theory: We see good performance for $\widehat{\Psi}_3$ and poor performance for the naive estimator.

A.6 Real Data Analysis: Network Recovery of Boston Housing Data

We evaluate our approach on the Boston housing data [47] by analyzing the network structure of features that may potentially impact house price. This dataset contains information

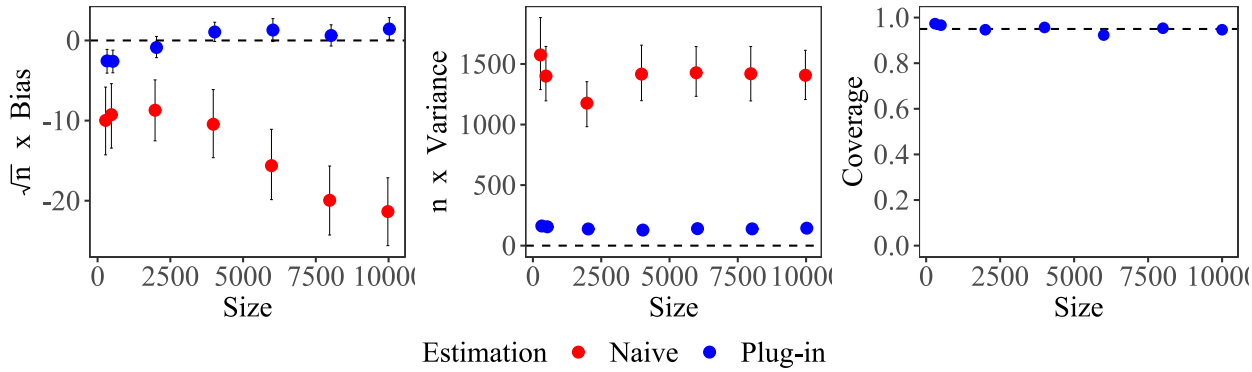

Figure A.3: Moderate dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_1(P)$. Conditional mean is estimated by gradient boosting.

collected by the U.S Census Service concerning housing in different areas of Boston Mass. There are 506 observations and each observation is based on a single town, with information on median home value (MEDV). In addition, it provides the four types of attributes which may be potential predictors to the price of house. The first type consists of neighborhood feature: % of lower socio-economic status (LSTAT); % of residential land zoned for lots larger than 25,000 square feet (ZN); % of black residents in the population (B); per capita crime rate by town (CRIM); % of non-retail business acres per town (INDUS); the full value property tax rate (TAX); the pupil-teacher ratio by school district (PTRATIO); Charles River dummy variable (CHAS). The second type is the house structural features: the average number of rooms per dwelling (RM) and % of owner-occupied units built prior to 1940 (AGE); The third one consists of accessibility features: index of accessibility to radial highways (RAD) and the weighted distances to five Boston employment centers (DIS). The final type is about air pollution, which only includes the nitric oxides concentration (NOX).

Here, we consider Gaussian graphical model (GMM) and the scaled expected conditional covariance Ψ_3 to build a network of 14 attributes. For Ψ_3, we estimate the conditional mean

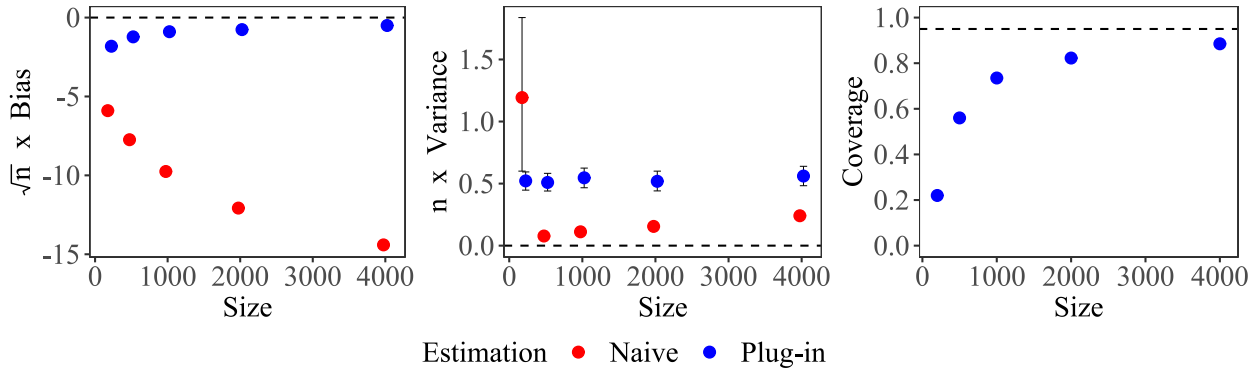

Figure A.4: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_3(P)$. Conditional mean is estimated by Lasso.

using random forests and we obtain p-values according to our asymptotic Gaussian limits as discussed in Theorem 2. For GMM, we use the bootstrap to build confidence intervals. In addition, we also use the value of $\widehat{\Psi}_3$ and the corresponding entry of the estimated precision matrix to represent the strength of association.

We display the results in Figure A.5. The network constructed by the scaled expected conditional covariance Ψ_3 shows that median house value (MEDV) is strongly connected with neighborhood and structural characteristics, such as the number of room (RM), weighted distances to employment centres (DIS), % of lower socio-economic status residents(LSTAT), crime rate (CRIM) and property-tax rate(TAX). This is similar to the findings of [13] and [123] where these attributes were also marked as important. In particularly, average number of room and proportion of lower socio-economic status, which were previously found as the most important feature, also have the strongest conditional association with the price.

In this example, estimating the network using GGM gives very different results. The graph structure is much less parsimonious. This is to be expected under model-misspecification: It is likely that the *true* precision matrix derived from complicated non-Gaussian data is

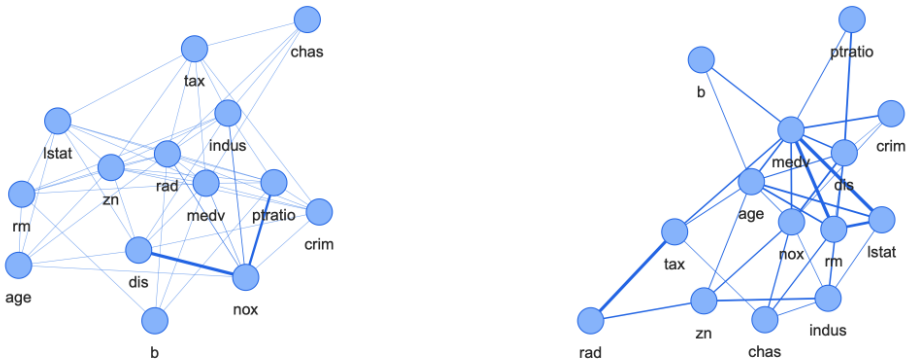

Figure A.5: Network constructed using GGM (left) and Ψ_3 (right). P-values are obtained to identify edges. Width of edges represents the corresponding entry in the precision matrix (left) or the value of Ψ_3 (right).

quite dense; it is unfortunately just a meaningless measure in such a case. In addition, edges connected to median price (MEDV) do not agree with previous published studies.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Proof of Proposition 3.2

If $X_j \sim \text{Bernoulli}(\mu_j)$, where $\mu_j := \text{E}(X_j|\mathbf{X}_{-j}) = p(X_j = 1|\mathbf{X}_{-j})$. Then,

$$\begin{aligned}
&\text{E}\{\text{Cov}(Y, X_j|\mathbf{X}_{-\mathbf{j}})\} \\
&= \text{E}\{\text{E}(YX_j|\mathbf{X}_{-j}) - \text{E}(Y|X_j)\,\text{E}(X_j|\mathbf{X}_{-j})\} \\
&= \text{E}\{\text{E}(Y|X_j = 1, \mathbf{X}_{-j})p(X_j = 1|\mathbf{X}_{-j}) - \text{E}(Y|\mathbf{X}_{-j})p(X_j = 1|\mathbf{X}_{-j})\} \\
&= \text{E}\{\text{E}(Y|X_j = 1, \mathbf{X}_{-j})p(X_j = 1|\mathbf{X}_{-j}) - p(X_j = 1|\mathbf{X}_{-j})\times \\
&\quad [\text{E}(Y|X_j = 1, \mathbf{X}_{-j})p(X_j = 1|\mathbf{X}_{-j}) + \text{E}(Y|X_j = 0, \mathbf{X}_{-j})p(X_j = 0|\mathbf{X}_{-j})]\} \\
&= \text{E}\{[\text{E}(Y|X_j = 1, \mathbf{X}_{-j}) - \text{E}(Y|X_j = 0, \mathbf{X}_{-j})]\mu_j(1 - \mu_j)\}
\end{aligned} \tag{B.1}$$

Therefore,

$$\Phi_j(P) = \text{E}\left\{\frac{\mu_j(1-\mu_j)}{\text{E}[\mu_j(1-\mu_j)]}\left(\text{E}[Y|X_j = 1, \mathbf{X}_{-j}] - \text{E}[Y|X_j = 0, \mathbf{X}_{-j}]\right)\right\}. \tag{B.2}$$

B.2 Proof of Lemma 3.1

For brevity, we let Φ_j denote $\Phi_j(P)$. Its estimator is $\widehat{\Phi}_j = \frac{\widehat{\Psi}_j}{\widehat{V}_j}$. So, the performance of $\widehat{\Phi}_j$ would be dependent on $\widehat{\Psi}_j$ and $\widehat{V}_j$. In Chapter 2, we have shown that

$$\begin{aligned}
\sqrt{n}\left(\widehat{\Psi}_j - \Psi_j(P)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{\Psi_j}(\mathbf{o}^{(i)}) + \Delta_j^{\Psi}, \\
\sqrt{n}\left(\widehat{V}_j - V_j(P)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_P^{V_j}(\mathbf{o}^{(i)}) + \Delta_j^{V}
\end{aligned} \tag{B.3}$$

where $D_P^{\Psi_j}(\mathbf{o}^{(i)}) = \left[y^{(i)} - \mu_{P,Y}(\mathbf{x}_{-j}^{(i)}\right] \left[x_j^{(i)} - \mu_{P,j}(\mathbf{x}_{-j}^{(i)}\right]$ and $D_P^{V_j}(\mathbf{o}^{(i)}) = \left[x_j^{(i)} - \mu_{P,j}(\mathbf{x}_{-j}^{(i)}\right]^2$ are the efficient influence functions of $\Psi_j(P)$ and $V_j(P)$. Thus, to prove Lemma 3.1, we calculate $\widehat{\Phi}_j - \Phi_j(P)$ as follows

$$
\begin{aligned}
\widehat{\Phi}_j - \Phi_j &= \frac{\widehat{\Psi}_j}{\widehat{V}_j} - \frac{\Psi_j}{V_j} \\
&= \frac{\widehat{\Psi}_j}{\widehat{V}_j} - \frac{\Psi_j}{\widehat{V}_j} + \frac{\Psi_j}{\widehat{V}_j} - \frac{\Psi_j}{V_j} \\
&= \frac{\widehat{\Psi}_j - \Psi_j}{V_j} - \frac{(\Psi_j - \widehat{\Psi}_j)(V_j - \widehat{V}_j)}{V_j \widehat{V}_j} + \frac{\Phi_j(V_j - \widehat{V}_j)}{V_j} + \frac{\Phi_j(V_j - \widehat{V}_j)^2}{V_j \widehat{V}_j} \\
&= \frac{\mathbb{P}_n D_P^{\Psi_j} + \Delta^{\Psi_j}/\sqrt{n}}{V_j} - \frac{\Phi_j(\mathbb{P}_n D_P^{V_j} + \Delta^{V_j}/\sqrt{n})}{V_j} + \frac{\Phi_j(V_j - \widehat{V}_j)^2}{V_j \widehat{V}_j} - \frac{(\widehat{\Psi} - \Psi)(\widehat{V}_j - V_j)}{V_j \widehat{V}_j} \\
&= \frac{\mathbb{P}_n D_P^{\Psi_j} - \Phi_j \mathbb{P}_n D_P^{V_j}}{V_j} + \frac{\Delta^{\Psi_j} - \Phi_j \Delta^{V_j}}{\sqrt{n} V_j} + \frac{\Phi_j(V_j - \widehat{V}_j)^2}{V_j \widehat{V}_j} - \frac{(\widehat{\Psi_j} - \Psi_j)(\widehat{V}_j - V_j)}{V_j \widehat{V}_j} \\
&= \mathbb{P}_n D_P^{\Phi_j} + \Delta^{\Phi_j}/\sqrt{n},
\end{aligned}
\tag{B.4}
$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}^{(i)})$ and $Pf = \int f dP$. Plug in $D_P^{\Psi_j}$ and $D_P^{V_j}$, we get that

$$
D_P^{\Phi_j}(\mathbf{o}) = \frac{\{y - \mu_{P,Y}(\mathbf{x}_{-j})\}\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}}{V_j} - \Phi_j \frac{\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}^2}{V_j}. \tag{B.5}
$$

which happens to be the gradient of the pathwise derivative of $\Phi_j(P)$ as we have shown in Chapter 2 for the parameter $\Psi_3(P)$ and its proof in Section 2.2, and thus is the efficient influence function of $\Phi_j(P)$.

Although, in the lemma, we assume a linear model. But the results still hold under the following partially linear model

$$
Y = X_j \beta_j + g(\mathbf{X}_{-j}) + \epsilon, \tag{B.6}
$$

where $X \sim (0, \Sigma_p)$ and $\epsilon_{i.i.d.} \sim (0, \sigma^2) \perp\!\!\!\perp X$. Then, $\Phi_j = \beta_j$ and the efficient influence

function $D_P^{\Phi_j}$ reduces to

$$\begin{aligned} & D_{P_{\text{lm}}}^{\Phi_j}(\mathbf{o}) \\ &= \frac{\{x_j\beta_j + g(\mathbf{x}_{-j}) + \epsilon - g(\mathbf{x}_{-j}) - \beta_j\mu_{P,j}(\mathbf{x}_{-j}) - x_j\beta_j + \beta_j\mu_{P,j}(\mathbf{x}_{-j})\}\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}}{V_j} \\ &= \frac{\{X_j - \mu_{P,j}(\mathbf{x}_{-j})\}\epsilon}{V_j}, \end{aligned} \tag{B.7}$$

as claimed.

B.3 Proof of Theorem 3.4

Following (B.4), the remainder term Δ^{Φ_j} has the following form

$$\Delta^{\Phi_j} = \frac{\Delta^{\Psi_j} - \Phi_j\Delta^{V_j}}{V_j} + \frac{\sqrt{n}\Phi_j(V_j - \widehat{V}_j)^2}{V_j\widehat{V}_j} - \frac{\sqrt{n}(\widehat{\Psi}_j - \Psi_j)(\widehat{V}_j - V_j)}{V_j\widehat{V}_j}, \tag{B.8}$$

where,

$$\begin{aligned} \Delta^{\Psi_j} &= \sqrt{n}(\mathbb{P}_n - P)\left[\widehat{D}^{\Psi_j}(\mathbf{o}) - D_P^{\Psi_j}(\mathbf{o})\right] + \sqrt{n}P[\widehat{\mu}_Y(\mathbf{x}_{-j}) - \mu_{P,Y}(\mathbf{x}_{-j})][\widehat{\mu}_j(\mathbf{x}_{-j}) - \mu_{P,j}(\mathbf{x}_{-j})] \\ \Delta^{V_j} &= \sqrt{n}(\mathbb{P}_n - P)\left[\widehat{D}^{V_j}(\mathbf{o}) - D_P^{V_j}(\mathbf{o})\right] + \sqrt{n}P[\widehat{\mu}_j(\mathbf{x}_{-j}) - \mu_{P,j}(\mathbf{x}_{-j})]^2, \end{aligned} \tag{B.9}$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}^{(i)})$ and $Pf = \int f dP$. $\widehat{D}^{\Psi_j}$ and $\widehat{D}^{V_j}$ are consistent estimator of $D_P^{\Psi_j}$ and $D_P^{V_j}$. The first summands in both Δ^{Ψ_j} and Δ^{V_j} are empirical process terms and can be shown to be asymptotically negligible by assuming the estimators belonging to P-Donsker class [115]. In addition, we also assume that the estimated conditional means converge to their truth, as in (3.16). Therefore,

$$\Delta^{\Psi_j} = o_P(1) \quad \text{and} \quad \Delta^{V_j} = o_P(1)$$

Meanwhile, when conditions in Theorem 3.4 are satisfied, Chapter 2 and [125] also have

shown that

$$\widehat{\Psi}_j - \Psi_j(P) = o_P(n^{-1/2}), \quad \widehat{V}_j - V_j(P) = o_P(n^{-1/2}). \tag{B.10}$$

Thus, the last 2 terms in (B.8) are $o_P(n^{-1/2})$. Therefore, Δ^{Φ_j} is negligible, i.e.,

$$\Delta^{\Phi_j} = o_P(1).$$

The performance of $\widehat{\Phi}_j$ will be dominated by $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_P^{\Phi_j}(\mathbf{o}^{(i)})$. By central limit theorem and Slutsky's theorem, we have

$$\sqrt{n}\left[\widehat{\Phi}_j - \Phi_j(P)\right] \to N\left(0, \sigma_j^2(P)\right),$$

where $\sigma_j^2(P) = \mathrm{Var}(D_P^{\Phi_j}(\mathbf{o})) = \int \left\{D_P^{\Phi_j}(\mathbf{o})\right\}^2 dP(\mathbf{o})$.

B.4 Proof of Theorem 3.5

Theorem 3.4 has told us the asymptotic normality of $\widehat{\Phi}_j$, i.e., $N\left(0, \sigma_j^2(P)\right)$. We can plug in the efficient influence function in (B.7) when linear model truly holds. There, the limiting variance becomes

$$\begin{aligned}
\sigma_j^2(P_{\mathrm{lm}}) &= \int \left\{D_{P_{\mathrm{lm}}}^{\Phi_j}(\mathbf{o})\right\}^2 dP(\mathbf{o}) \\
&= \int \left\{\frac{\{x_j - \mu_{P,j}(\mathbf{x}_{-j})\}\,\epsilon}{V_j}\right\}^2 dP(\mathbf{o}) \\
&= \sigma^2 / V_j.
\end{aligned}$$

If we further assume that $\mathbf{X} \sim N(0, \Sigma_p)$, then

$$\begin{aligned}
V_j = \mathrm{E}[\mathrm{Var}(X_j|\mathbf{X}_{-j})] &= \mathrm{Var}(X_j|\mathbf{X}_{-j}) \\
&= \Sigma_p(j,j) - \Sigma_p(j,-j)\Sigma_p^{-1}(-j,-j)\Sigma_p(-j,j) \\
&= 1/\Theta_{p,jj}
\end{aligned} \tag{B.11}$$

Hence, the limiting distribution of $\widehat{\Phi}_j$ is

$$\sqrt{n}\left[\widehat{\Phi}_j - \beta_j\right] \to_d N\left(0, \sigma^2 \Theta_{p,jj}\right) \tag{B.12}$$

which is the same as the limiting distribution of the OLS estimator $\widehat{\beta}_j^{\text{OLS}}$.

B.5 Proof of Lemma 3.2

For ease of the exposition, we write $\Phi_{n,j} := \Phi_j(P_n)$, $\Psi_{n,j} := \Psi_j(P_n)$, and $V_{n,j} := V_j(P_n)$. As shown in Proposition 3.3,

$$\sqrt{n}\left(\widehat{\Phi}_{n,j} - \Phi_{n,j}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n^{(i)}) + \Delta_n^{\Phi_j}. \tag{B.13}$$

The remainder $\Delta_n^{\Phi_j}$ can be expressed as

$$\Delta_n^{\Phi_j} = \frac{\Delta_n^{\Psi_j} - \Phi_{n,j}\Delta_n^{V_j}}{V_{n,j}} + \frac{\sqrt{n}\Phi_{n,j}(V_{n,j} - \widehat{V}_{n,j})^2}{V_{n,j}\widehat{V}_{n,j}} - \frac{\sqrt{n}(\widehat{\Psi}_{n,j} - \Psi_{n,j})(\widehat{V}_{n,j} - V_{n,j})}{V_{n,j}\widehat{V}_{n,j}}. \tag{B.14}$$

Just like in the low-dimensional case, we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\Psi}_{n,j} - \Psi_{n,j}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{P_n}^{\Psi_j}(\mathbf{o}^{(i)}) + \Delta_n^{\Psi_j}, \\
\sqrt{n}\left(\widehat{V}_{n,j} - V_{n,j}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{P_n}^{V_j}(\mathbf{o}^{(i)}) + \Delta_n^{V_j}
\end{aligned} \tag{B.15}$$

and

$$\begin{aligned}
\Delta_n^{\Psi_j} &= \sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}_n^{\Psi_j}(\mathbf{o}_n) - D_{P_n}^{\Psi_j}(\mathbf{o}_n)\right] + \\
&\qquad \sqrt{n}P_n[\widehat{\mu}_{n,Y}(\mathbf{x}_{n,-j}) - \mu_{P_n,Y}(\mathbf{x}_{n,-j})][\widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}) - \mu_{P_n,j}(\mathbf{x}_{n,-j})] \\
\Delta_n^{V_j} &= \sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}^{V_j}(\mathbf{o}_n) - D_P^{V_j}(\mathbf{o}_n)\right] + \sqrt{n}P_n[\widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}) - \mu_{P_n,j}(\mathbf{x}_{n,-j})]^2.
\end{aligned} \tag{B.16}$$

By assuming that,

$$\int [\widehat{\mu}_{n,Y}(\mathbf{x}_{n,-j}) - \mu_{P_n,Y}(\mathbf{x}_{n,-j})]^2 dP_n(\mathbf{x}_{n,-j}) = o_{P_n}\left(n^{-1/2}\right),$$
$$\int [\widehat{\mu}_{n,j}(\mathbf{x}_{n,-j}) - \mu_{P_n,j}(\mathbf{x}_{n,-j})]^2 dP_n(\mathbf{x}_{n,-j}) = o_{P_n}\left(n^{-1/2}\right),$$

and the empirical process terms can be bounded as n and p_n go to infinity:

$$\sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}_n^{\Psi_j}(\mathbf{o}_n) - D_{P_n}^{\Psi_j}(\mathbf{o}_n)\right] = o_{P_n}(1),$$
$$\sqrt{n}(\mathbb{P}_n - P_n)\left[\widehat{D}_n^{V_j}(\mathbf{o}_n) - D_{P_n}^{V_j}(\mathbf{o})\right] = o_{P_n}(1).$$

Then, we get that $\Delta_n^{\Psi_j} = o_{P_n}(1)$ and $\Delta_n^{V_j} = o_{P_n}(1)$, followed by $\frac{\Delta_n^{\Psi_j} - \Phi_{n,j}\Delta_n^{V_j}}{V_{n,j}} = o_{P_n}(1)$.

Meanwhile, we can use the same swapping and interpolation techniques in the following section, i.e., Section B.6, to show that $\widehat{\Psi}_{n,j} - \Psi_{n,j} = o_{P_n}(n^{-1/2})$ and $\widehat{V}_{n,j} - V_{n,j} = o_{P_n}(n^{-1/2})$. Thus, we have $\Delta_{P_n}^{\Phi_j} = o_{P_n}(1)$.

B.6 Proof of Theorem 3.7

We first prove Theorem 3.7, where we use the techniques of swapping and interpolation that can also be applied for Theorem 3.6, as well as the proof for Lemma 3.2. As shown in Proposition 3.3,

$$\sqrt{n}\left(\widehat{\Phi}_{n,j} - \beta_{n,j}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n^{(i)}) + \Delta_n^{\Phi_j}. \tag{B.17}$$

where $D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n) = \frac{\{y_n - \mu_{P_n,Y}(\mathbf{x}_{n,-j})\}\{x_{n,j} - \mu_{P_n,j}(\mathbf{x}_{n,-j})\}}{V_{n,j}} - \Phi_{n,j}\frac{\{x_{n,j} - \mu_{P_n,j}(\mathbf{x}_{n,-j})\}^2}{V_{n,j}}$ and $\Delta_n^{\Phi_j} = o_{P_n}(1)$ by Lemma 3.2. Let

$$w_n^{(i)} = \frac{D_{P_n}^{\Phi_{n,j}}(\mathbf{o}_n^{(i)})}{\sigma_{n,j}} \tag{B.18}$$

where $\sigma_{n,j}^2 := \sigma_j^2(P_n) = \int \{D_{P_n}^{\Phi_{n,j}}\}^2 dP_n$. Thus, $w_n^{(i)} \sim_{i.i.d.} w_n$ with $\mathrm{E}(w_n) = 0$ and $\mathrm{Var}(w_n) = 1$. Let $W_n = \left(w_n^{(1)} + w_n^{(2)} + \ldots + w_n^{(n)}\right)/\sqrt{n}$ and $Z_n = \left(z_n^{(1)} + z_n^{(2)} + \ldots + z_n^{(n)}\right)/\sqrt{n}$ with

$z_n^{(i)} \sim_{iid} N(0,1)$. Then, we can use the swapping technique [25] to show the convergence of W_n. Let

$$\begin{aligned} V_n^{(i)} &= \left(w_n^{(1)} + \ldots + w_n^{(i)} + z_n^{(i+1)} + \ldots + z_n^{(n)}\right)/\sqrt{n} \\ U_n^{(i)} &= \left(w_n^{(1)} + \ldots + w_n^{(i-1)} + z_n^{(i+1)} + \ldots + z_n^{(n)}\right)/\sqrt{n} \end{aligned} \tag{B.19}$$

with $V_n^{(0)} = Z_n$ and $V_n^{(n)} = W_n$ such that

$$\begin{aligned} V_n^{(i)} &= U_n^{(i)} + w_n^{(i)}/\sqrt{n} \\ V_n^{(i-1)} &= U_n^{(i)} + z_n^{(i)}/\sqrt{n} \end{aligned} \tag{B.20}$$

Consider a smooth function $h(x)$ which has bounded third-order derivative. By Taylor expansion, we get that:

$$\begin{aligned} h(U_n^{(i)} + w_n^{(i)}/\sqrt{n}) = h(U_n^{(i)}) + w_n^{(i)}/\sqrt{n}h'(U_n^{(i)}) + \frac{1}{2n}{w_n^{(i)}}^2 h''(U_n^{(i)}) \\ + \frac{1}{6}n^{-3/2}{w_n^{(i)}}^3 h'''(U_n^{(i)} + u \times w_n^{(i)}/\sqrt{n}) \end{aligned} \tag{B.21}$$

$$\begin{aligned} h(U_n^{(i)} + z_n^{(i)}/\sqrt{n}) = h(U_n^{(i)}) + z_n^{(i)}/\sqrt{n}h'(U_n^{(i)}) + \frac{1}{2n}{z_n^{(i)}}^2 h''(U_n^{(i)}) \\ + \frac{1}{6}n^{-3/2}{z_n^{(i)}}^3 h'''(U_n^{(i)} + v \times z_n^{(i)}/\sqrt{n}) \end{aligned} \tag{B.22}$$

for some $u, v \in [0,1]$. Then

$$\begin{aligned} & h(V_n^{(i)}) - h(V_n^{(i-1)}) \\ &= h(U_n^{(i)} + w_n^{(i)}/\sqrt{n}) - h(U_n^{(i)} + z_n^{(i)}/\sqrt{n}) \\ &= \left(w_n^{(i)} - z_n^{(i)}\right)/\sqrt{n}h'(U_n^{(i)}) + \frac{1}{2n}\left({w_n^{(i)}}^2 - {z_n^{(i)}}^2\right)[h''(U_n^{(i)})]^2 \\ &\quad + \frac{1}{6}n^{-3/2}{w_n^{(i)}}^3 h'''(U_n^{(i)} + u \times w_n^{(i)}/\sqrt{n}) - \frac{1}{6}n^{-3/2}{z_n^{(i)}}^3 h'''(U_n^{(i)} + v \times z_n^{(i)}/\sqrt{n}) \end{aligned} \tag{B.23}$$

Taking the expectation on the both sides of the above equation and using the fact that

$w_n^{(i)}$ and $z_n^{(i)}$ are independent of $U_n^{(i)}$, it follows that

$$\begin{aligned}
&\mathrm{E}\, h(W_n) - \mathrm{E}\, h(Z_n) = \mathrm{E} \sum_{i=1}^{n} \left[h(V_n^{(i)}) - h(V_n^{(i-1)}) \right] \\
&= \frac{1}{6} n^{-3/2} \sum_{i=1}^{n} \mathrm{E}\, {w_n^{(i)}}^3 h'''(U_n^{(i)} + u w_n^{(i)}/\sqrt{n}) - \frac{1}{6} n^{-3/2} \sum_{i=1}^{n} \mathrm{E}\, {z_n^{(i)}}^3 h'''(U_n^{(i)} + v z_n^{(i)}/\sqrt{n})
\end{aligned} \tag{B.24}$$

For $P(W_n \le z) = \mathrm{E}\, 1[W_n \le z]$ and a tiny $c > 0$, we consider a smooth function $h_c(x)$ which (i) equals to 1 on $(-\infty, z)$; (ii) vanishes outside of $[z + c, \infty)$; and (iii) has third order derivative $h_c'''(x) = O(c^{-3})$. Let $Z \sim N(0,1)$, then $z_n^{(i)} \equiv Z$ and thus

$$\begin{aligned}
P(W_n \le z) - P(Z \le z) &= \mathrm{E}\, 1[W_n \le z] - \mathrm{E}\, h_c(Z_n) + \mathrm{E}\, h_c(Z_n) - P(Z \le z) \\
&\le \mathrm{E}\, h_c(W_n) - \mathrm{E}\, h_c(Z_n) + \mathrm{E}\, h_c(Z_n) - P(Z \le z) \\
&= \frac{1}{6} n^{-3/2} \sum_{i=1}^{n} \mathrm{E}\, {w_n^{(i)}}^3 h_c'''(U_n^{(i)} + u w_n^{(i)}/\sqrt{n}) \\
&\quad - \frac{1}{6} n^{-3/2} \sum_{i=1}^{n} \mathrm{E}\, {z_n^{(i)}}^3 h_c'''(U_n^{(i)} + v z_n^{(i)}/\sqrt{n}) + P(z \le z_n^{(i)} \le z + c) \\
&\le \frac{1}{6c^3} \sum_{i=1}^{n} \mathrm{E} \left| w_n^{(i)}/\sqrt{n} \right|^3 + c/\sqrt{2\pi} \\
&= \frac{n^{-1/2}\, \mathrm{E}\, |w_n|^3}{6c^3} + c/\sqrt{2\pi} \\
\xrightarrow{c = \left(\sqrt{2\pi} n^{-1/2}\, \mathrm{E}\, |w_n|^3/6 \right)^{1/4}} &\le (C\, \mathrm{E}\, |w_n|^3/\sqrt{n})^{1/4}
\end{aligned} \tag{B.25}$$

for some constant $C > 0$. Hence, we finally have

$$P(W_n \le z) - P(Z \le z) = O\left[(\mathrm{E}\, |w_n|^3/\sqrt{n})^{1/4} \right] = o(1) \tag{B.26}$$

If the linear model (3.24) holds with $\mathbf{x}_n^{(i)} \sim_{i.i.d.} N(0, \Sigma_{p_n})$, then $\sigma_j^2(P_n) = \sigma^2 / V_j(P_n) =$

$\sigma^2\Theta_{p_n,jj}$. Then, it is natural to get

$$\frac{\sqrt{n}(\widehat{\Phi}_{n,j}-\beta_{n,j})}{\sigma\Theta_{p_n,jj}}=W_n+o_{P_n}(1). \tag{B.27}$$

Since, the influence function under linearity becomes $D^{\Phi_{n,j}}_{P_{n,\text{lm}}}(\mathbf{o}_n)=\frac{[x_{n,j}-\mu_{P_n,j}(\mathbf{x}_{n,-j})]\epsilon_n}{V_j(P_n)}$. So, $w_n^{(i)}$ reduces to $w_n^{(i)}=\frac{[x_{n,j}-\mu_{P_n,j}(\mathbf{x}_{n,-j})]\epsilon_n}{\sigma^2}$. Using the same swapping techniques, we can arrive at the conclusion in (B.26).

B.7 Proof of Theorem 3.6

In the triangular array setup, the de-biased lasso estimator is

$$\widehat{\boldsymbol{\beta}}_n^{DL}=\widehat{\boldsymbol{\beta}}_n^{Lasso}+\widehat{\Theta}_{p_n}\mathbf{x}_n^T(\mathbf{y}_n-\mathbf{x}_n\widehat{\boldsymbol{\beta}}_n^{Lasso})/n, \tag{B.28}$$

where $\widehat{\Theta}_{p_n}$ is the node-wise lasso estimator in (3.28). When the linear model (3.24) truly holds, the difference between $\widehat{\boldsymbol{\beta}}_n^{DL}$ and the truth $\boldsymbol{\beta}_n$ is

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{DL}-\boldsymbol{\beta}_n)&=\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{Lasso}+\widehat{\Theta}_{p_n}\mathbf{x}_n^T(\mathbf{x}_n\boldsymbol{\beta}_n+\boldsymbol{\epsilon}_n-\mathbf{x}_n\widehat{\boldsymbol{\beta}}_n^{Lasso})/n-\boldsymbol{\beta}_n)\\&=\sqrt{n}(I_{p_n}-\widehat{\Theta}_{p_n}\widehat{\Sigma}_{p_n})(\widehat{\boldsymbol{\beta}}_n^{Lasso}-\boldsymbol{\beta}_n)+\widehat{\Theta}_{p_n}\mathbf{x}_n^T\boldsymbol{\epsilon}_n/\sqrt{n}\\&=\Theta_{p_n}\mathbf{x}_n^T\boldsymbol{\epsilon}_n/\sqrt{n}+\sqrt{n}(I_{p_n}-\widehat{\Theta}_{p_n}\widehat{\Sigma}_{p_n})(\widehat{\boldsymbol{\beta}}_n^{Lasso}-\boldsymbol{\beta}_n)+(\widehat{\Theta}_{p_n}-\Theta_{p_n})\mathbf{x}_n^T\boldsymbol{\epsilon}_n/\sqrt{n}\\&=\Theta_{p_n}\mathbf{x}_n^T\boldsymbol{\epsilon}_n/\sqrt{n}+\Delta_n^{\text{DL}}.\end{aligned} \tag{B.29}$$

Therefore, to estimate the jth coefficient, we have

$$\frac{\sqrt{n}(\widehat{\beta}_{n,j}^{DL}-\beta_{n,j})}{\sigma\sqrt{\Theta_{p_n,jj}}}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}r_n^{(i)}+\frac{\Delta_{n,j}^{\text{DL}}}{\sigma\sqrt{\Theta_{p_n,jj}}} \tag{B.30}$$

where $r_n^{(i)}=\frac{\Theta_{p_n,j}^T\mathbf{x}_n^{(i)}\epsilon_n^{(i)}}{\sigma\sqrt{\Theta_{p_n,jj}}}$. $r_n^{(i)}\sim_{ind}(0,1)$, i.e., we say that $r_n^{(i)}\equiv r_n$ with $\text{E}(r_n)=0$ and $\text{Var}(r_n)=1$. Let $R_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}r_n^{(i)}$, we can also use the same swapping and interpolation

technique as in Section B.6 to show that

$$\mathbb{P}\left(R_n \geq z\right) - \mathbb{P}(Z > z) = O\left[\left(\mathrm{E}\left|r_n\right|^3/\sqrt{n}\right)^{1/4}\right] = o(1) \tag{B.31}$$

if $\mathrm{E}\left|r_n^3\right| = O(1)$. Z is a standard normal variable.

By Assumption 3.2, $\|\Theta_{p_n}\|_\infty = \mathcal{O}(1)$. Then, the rest of this proof is to show that $\Delta_{n,j}^{\mathrm{DL}} = o_{P_n}(1)$. The Karush–Kuhn–Tucker (KKT) condition for the node-wise lasso regression (3.28) implies that

$$\lambda_j \widehat{\kappa}_{n,j} = \mathbf{x}_{n,-j}^T(\mathbf{x}_{n,j} - \mathbf{x}_{n,-j}\widehat{\boldsymbol{\gamma}}_{n,j})/n \tag{B.32}$$

where $\widehat{\kappa}_{n,j} = \mathrm{sign}(\widehat{\boldsymbol{\gamma}}_{n,j})$. By $\|\widehat{\boldsymbol{\gamma}}_{n,j}\|_1 = \widehat{\boldsymbol{\gamma}}_{n,j}^T\widehat{\kappa}_{n,j}$, we get

$$\widehat{\tau}_{n,j}^2 = \mathbf{x}_{n,j}^T(\mathbf{x}_{n,j} - \mathbf{x}_{n,-j}\widehat{\boldsymbol{\gamma}}_{n,j})/n. \tag{B.33}$$

Recall that $\widehat{\Theta}_{p_n,j} = \widehat{\mathcal{C}}_{n,j}/\widehat{\tau}_{n,j}^2$, implying that

$$\begin{aligned} \mathbf{x}_{n,j}^T\mathbf{x}_n\widehat{\Theta}_{p_n,j}/n &= \mathbf{x}_{n,j}^T\mathbf{x}_n\widehat{\mathcal{C}}_{n,j}/(n\widehat{\tau}_{n,j}^2) \\ &= \mathbf{x}_{n,j}^T(\mathbf{x}_{n,j} - \mathbf{x}_{n,-j}\widehat{\boldsymbol{\gamma}}_{n,j})/(n\widehat{\tau}_{n,j}^2) \\ &= 1. \end{aligned} \tag{B.34}$$

Meanwhile, followed by (B.32),

$$\|\mathbf{x}_{n,-j}^T\mathbf{x}_n\widehat{\Theta}_{p_n,j}\|_\infty/n = \|\mathbf{x}_{n,-j}^T\mathbf{x}_n\widehat{\mathcal{C}}_{n,j}/\widehat{\tau}_{n,j}^2\|_\infty/n = \lambda_j\|\widehat{\kappa}_{n,j}/\widehat{\tau}_{n,j}^2\|_\infty \leq \lambda_j/\widehat{\tau}_{n,j}^2. \tag{B.35}$$

Therefore,

$$\|\widehat{\Sigma}_{p_n}\widehat{\Theta}_{p_n,j} - e_j\|_\infty = \|(\mathbf{x}_{n,j}, \mathbf{x}_{n,-j})^T\mathbf{x}_n\widehat{\Theta}_{p_n,j}/n - e_j\|_\infty \leq \lambda_j/\widehat{\tau}_{n,j}^2, \tag{B.36}$$

and thus $\|\widehat{\Sigma}_{p_n}\widehat{\Theta}_{p_n} - I_{p_n}\|_\infty \leq \max_j \lambda_j/\widehat{\tau}_{n,j}^2$.

The following proposition tells that under some assumptions, $\lambda_j/\widehat{\tau}_{n,j}^2$ can be uniformly

bounded, for all $j = 1, ..., p_n$.

Lemma 5.3 in this work [111] explicitly shows that, under our Assumption 3.2, if the sparsity with respect to the columns of the precision matrix Θ_{p_n},is bounded by

$$\max_{1 \le j \le p_n} s_{n,j} = o(n/\log(p_n)), \tag{B.37}$$

and for each $j = 1, ..., p_n$, the regularization parameter in the node-wise lasso regression is suitably chosen, that is $\lambda_j \asymp \sqrt{\log(p_n)/n}$. Then,

$$\max_j 1/\widehat{\tau}^2_{n,j} = O_{P_n}(1), \tag{B.38}$$

where we say $X_n = O_{P_n}(1)$ if $P_n(|X_n| < \delta_c) \ge 1 - c$, $\forall c > 0$ and $n \ge N$. Then,

$$\|\widehat{\Sigma}_{p_n}\widehat{\Theta}_{p_n} - I_{p_n}\|_\infty \le \max_j \lambda_j/\widehat{\tau}^2_{n,j} \asymp \sqrt{\log(p_n)/n}. \tag{B.39}$$

Meanwhile, by additionally making the sparsity assumption of the model, i.e., $s_{n,0} = o(\sqrt{n}/\log(p_n))$, and suitably selecting the regularization parameter in (3.26), with $\lambda \asymp \sqrt{\log(p_n)/n}$, Lemma 5.1 and 5.2 in [111] also provide that

$$\|\widehat{\boldsymbol{\beta}}^{\text{Lasso}}_n - \boldsymbol{\beta}_n\|_1 = O_{P_n}\left(s_{n,0}\sqrt{\log(p_n)/n}\right) = o_{P_n}(1/\sqrt{\log(p_n)}). \tag{B.40}$$

Therefore, we can bound the first term in $\Delta^{\text{DL}}_{n,j}$ as

$$\begin{aligned}\|\sqrt{n}(I_{p_n} - \widehat{\Theta}_{p_n}\widehat{\Sigma}_{p_n})(\widehat{\boldsymbol{\beta}}^{Lasso}_n - \boldsymbol{\beta}_n)\|_\infty &\le \sqrt{n}\|I_{p_n} - \widehat{\Theta}_{p_n}\widehat{\Sigma}_{p_n}\|_\infty\|\widehat{\boldsymbol{\beta}}^{Lasso}_n - \boldsymbol{\beta}_n\|_1 \\ &= \sqrt{n}o_{P_n}(1/\sqrt{n}) = o_{P_n}(1).\end{aligned} \tag{B.41}$$

To bound $(\widehat{\Theta}_{p_n} - \Theta_{p_n})\mathbf{x}^T_n\boldsymbol{\epsilon}_n/\sqrt{n}$, we have

$$\|(\widehat{\Theta}_{p_n} - \Theta_{p_n})\mathbf{x}^T_n\boldsymbol{\epsilon}_n/\sqrt{n}\|_\infty \le \max_j |\mathbf{x}^T_{n,j}\boldsymbol{\epsilon}_n/\sqrt{n}| \times \|\widehat{\Theta}_{p_n} - \Theta_{p_n}\|_\infty. \tag{B.42}$$

Using the standard arguments for the $\mathcal{L}_2$-norm bound, i.e.,

$$\|\widehat{\Theta}_{p_n} - \Theta_{p_n}\|_\infty \leq \max_j \|\widehat{\Theta}_{p_n,j} - \Theta_{p_n}\|_2 \leq \max_j \lambda_j \sqrt{s_{n,j}} = o_{P_n}(1). \tag{B.43}$$

In addition, $x_{n,j}^{(i)}$ and $\epsilon_n^{(i)}$ are both Gaussian variables and are independent with each other. For any $j = 1, ..., p_n$, $x_{n,j}^{(i)} \epsilon_n^{(i)} / \sqrt{\sigma^2 \Sigma_{p_n,jj}} \sim_{ind} N(0,1)$. Since $\|\Sigma_{p_n}\|_\infty = O(1)$, $\mathbf{x}_{n,j}^T \boldsymbol{\epsilon}_n / \sqrt{n} = O_{P_n}(1)$, uniformly for $j = 1, .., p_n$. Thus,

$$\|(\widehat{\Theta}_{p_n} - \Theta_{p_n}) \mathbf{x}_n^T \boldsymbol{\epsilon}_n / \sqrt{n}\|_\infty = o_{P_n}(1). \tag{B.44}$$

This completes our proof of showing $\Delta_{n,j}^{\mathrm{DL}} = o_{P_n}(1)$.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Proof of Lemma 4.3

Recall that M is $\mathcal{F}$-embeddable and $\mathcal{F}$ satisfies Condition 4.1. Thus, the entries of M are generated by $m_{ij} = f_j(\boldsymbol{\theta}_{i,\cdot})$. Consider arbitrary $\epsilon > 0$. Then there is some fixed $C_0 > 0$, and a collection of functions $\mathcal{F}_\epsilon^* = \left\{\widetilde{\psi}_1, \widetilde{\psi}_2, \ldots, \widetilde{\psi}_{J^*(\epsilon)}\right\} \subset \mathcal{F}$ that give the finite set of minimal cardinality $J^*(\epsilon)$, with the property that $\max_{f\in\mathcal{F}} \min_{\|\beta\|_2^2 \leq C_0} \left\| f - \sum_{l=1}^{J^*(\epsilon)} \beta_l \widetilde{\psi}_l \right\|_\infty \leq \epsilon$. and $\|\widetilde{\psi}_l\|_\infty \leq C_0$. For any given $f \in \mathcal{F}$, let

$$\beta^\epsilon(f) = \text{argmin}_\beta \left\| f - \sum_{\widetilde{\psi}_l \in \mathcal{F}_\epsilon^*} \beta_l \widetilde{\psi}_l \right\|_\infty .$$

This implies that we can approximate M with a low rank matrix M^ϵ, with entries given by

$$m_{ij}^\epsilon \leftarrow \sum_{\widetilde{\psi}_l \in \mathcal{F}_\epsilon^*} \widetilde{\psi}_l\left(\boldsymbol{\theta}_{i,\cdot}\right) \cdot \beta_l^\epsilon(f_j), \tag{C.1}$$

such that $|m_{ij} - m_{ij}^\epsilon| \leq \epsilon$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$. Now, let Ψ denote the matrix with $\Psi_{il} = \widetilde{\psi}_l\left(\boldsymbol{\theta}_{i,\cdot}\right)$ and B denote the matrix with entries $B_{lj} = \beta_l^\epsilon(f_j)$. Then, the approximation matrix can be compactly written as

$$M^\epsilon = \Psi B,$$

with $\Psi \in \mathbb{R}^{n\times J^*(\epsilon)}$ and $B \in \mathbb{R}^{J^*(\epsilon)\times p}$. Thus, $\text{rank}(M^\epsilon) = J^*(\epsilon) \leq \min(n, p)$ and

$$\|M^\epsilon - M\|_\infty \leq \epsilon.$$

Finally, using a variational form of the nuclear norm [100], we have

$$\frac{1}{\sqrt{np}}\left\|M^{\epsilon}\right\|_{*}=\frac{1}{2}\min_{UV^{\top}=M^{\epsilon}}\left(\frac{1}{n}\left\|U\right\|_{F}^{2}+\frac{1}{p}\left\|V\right\|_{F}^{2}\right).$$

From the above statement, we know that $\|\Psi\|_{\infty}$ and $\|B\|_{\infty}$ are both bounded by C_0.

Thus we have that

$$\frac{1}{\sqrt{np}}\left\|M^{\epsilon}\right\|_{*}\leq\frac{1}{2}\left(\frac{1}{n}\left\|\Psi\right\|_{F}^{2}+\frac{1}{p}\left\|B\right\|_{F}^{2}\right)\leq C_{0}^{2}J^{*}(\epsilon).$$

Noting that C_0^2 is a constant independent of ϵ gives us our result.

C.2 Deriving the Consistency

In this section, we shall derive the consistency of our estimator $\widehat{M}$. Recall that $\{(y_t, X_t)\}_{t=1}^{N}$ are generated by

$$y_t = \langle X_t, M\rangle + \xi_t, \tag{C.2}$$

where ξ_t are i.i.d random errors distributed $N(0, \sigma^2)$, and M is a $n \times p$ matrix. The estimator we consider is defined by

$$\begin{aligned}\widehat{M} &\leftarrow \operatorname{argmin}_{M\in\mathbb{R}^{n\times p}}\left\{\frac{1}{np}\|M\|_F^2-\left\langle\frac{2}{N}\sum_{t=1}^{N}y_tX_t, M\right\rangle+\lambda\|M\|_*\right\}\\ &\equiv \operatorname{argmin}_{M\in\mathbb{R}^{n\times p}} L_N(M)\end{aligned} \tag{C.3}$$

We first introduce two technical lemmas, which will play the key role in showing the convergence rate. Proving these lemmas will entail most of the work required for proving this theorem. In Lemma 3.5, we derive a deterministic upper bound for the estimation error (under a stochastic condition) as a function of the regularization parameter λ when λ is sufficiently large (in this Lemma, "sufficiently large" is left as a stochastic constraint). In particular, we show that the risk can be decomposed into a misspecification error and a prediction error. Then, in Lemma 3.6, we identify a deterministic value for λ such that, with

high probability, the condition in Lemma 3.5 will hold. More specifically we give probabilistic bounds for the operator norm of the stochastic error term in our generative model. We can then combine these to obtain the general oracle inequality in Theorem 4.8.

Before continuing, we give some additional notation: For any matrix Z, we denote $\|Z\|_{op} = \Lambda_{\max}(Z)$, where $\Lambda^2_{\max}(Z) = \Lambda_{\max}(Z^T Z)$ is the largest singular value of $Z^T Z$, also known as the operator-norm.

Lemma 3.5. *Suppose we observe $\{(y_t, X_t)\}_{t=1}^N$ generated by (C.2), where X_t are i.i.d uniformly sampled from $\mathcal{X}$. Further, assume the underlying true matrix $M \in \mathbb{R}^{n\times p}$ is $\mathcal{F}$-embeddable with Condition 4.1 satisfied. Let $\Delta = N^{-1}\sum_{t=1}^N [y_t X_t - \mathrm{E}(y_t X_t)]$. If $\lambda \geq 2\|\Delta\|_{op}$, then*

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 \leq \epsilon^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 J^*(\epsilon)\lambda^2 np \tag{C.4}$$

holds for any $\epsilon > 0$. Recall that $J^*(\epsilon)$ is the minimal rank of an approximation matrix M^ϵ with $\|M^\epsilon - M\|_\infty < \epsilon$.

Proof. The proof of this lemma is based on the strong convexity of the loss function $L_N(M)$.

Consider the the subdifferential of $L_N(M)$, which is the set of matrices of the following form:

$$\partial L_N(M) = \left\{ \frac{2}{np} M - \frac{2}{N}\sum_{t=1}^N y_t X_t + \lambda B, \quad B \in \partial\|M\|_* \right\}. \tag{C.5}$$

Thus, the following representation holds for $\widehat{A} \in \partial L_N(\widehat{M})$

$$\widehat{A} = \frac{2}{np}\widehat{M} - \frac{2}{N}\sum_{t=1}^N y_t X_t + \lambda \widehat{B},$$

for some $\widehat{B} \in \partial\|\widehat{M}\|_*$. Since $M \mapsto L_N(M)$ is strictly convex, $\widehat{M}$ defined in (C.3) is the unique minimizer of $L_N(M)$. This implies, $\mathbf{0} \in \partial L_N(\widehat{M})$. Hence, there exists $\widehat{B} \in \partial\|\widehat{M}\|_*$ such that $\widehat{A} = \mathbf{0}$, and thus

$$\langle \widehat{A}, \widehat{M} - M^\epsilon \rangle = \langle \mathbf{0}, \widehat{M} - M^\epsilon \rangle = 0. \tag{C.6}$$

It further follows that

$$\begin{aligned}&\langle \widehat{A}, \widehat{M} - M^\epsilon \rangle \\ &= \frac{2}{np}\langle \widehat{M}, \widehat{M} - M^\epsilon \rangle - \frac{2}{N}\sum_{t=1}^{N}\langle y_t X_t, \widehat{M} - M^\epsilon \rangle + \lambda \langle \widehat{B}, \widehat{M} - M^\epsilon \rangle = 0.\end{aligned} \tag{C.7}$$

$M^\epsilon \in \mathbb{R}^{n\times p}$ is the approximation matrix with $\text{rank}(M^\epsilon) = J^*(\epsilon)$. So, it has spectral representation $M^\epsilon = \sum_{j=1}^{J^*(\epsilon)} \sigma_j u_j v_j^T$ where $u_j \in \mathbb{R}^n$ and $v_j \in \mathbb{R}^p$, $j = 1, ..., J^*(\epsilon)$, are orthonormal vectors, and σ_j are the singular values of M^ϵ. Let U and V denote the linear span of $\{u_1, ..., u_{J^*(\epsilon)}\}$ and $\{v_1, ..., v_{J^*(\epsilon)}\}$ respectively. Then, the subdifferential of $\|M^\epsilon\|_*$ can be represented by the following set of matrices [119]:

$$\partial \|M^\epsilon\|_* = \left\{ \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T + P_{U^\perp} W P_{V^\perp} : \|W\|_{op} \leq 1 \right\},$$

where $U^\perp$ denotes the orthogonal complements of U and $P_{U^\perp}$ denotes the projection on the linear vector subspace $U^\perp$. The same argument applies to V and $P_{V^\perp}$. Thus, $B^\epsilon \in \partial \|M^\epsilon\|_*$ can be represented as

$$B^\epsilon = \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T + P_{U^\perp} W P_{V^\perp} \tag{C.8}$$

for arbitrary matrix W having $\|W\|_{op} \leq 1$. Due to the trace duality, there exists W with $\|W\|_{op} \leq 1$ such that

$$\langle P_{U^\perp} W P_{V^\perp}, \widehat{M} - M^\epsilon \rangle = \langle P_{U^\perp} W P_{V^\perp}, \widehat{M} \rangle = \langle W, P_{U^\perp} \widehat{M} P_{V^\perp} \rangle = \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_* \tag{C.9}$$

So, it follows from (C.7) that

$$\begin{aligned}&\frac{2}{np}\langle \widehat{M} - M, \widehat{M} - M^\epsilon \rangle + \frac{2}{np}\langle M, \widehat{M} - M^\epsilon \rangle + \lambda \langle \widehat{B} - B^\epsilon, \widehat{M} - M^\epsilon \rangle \\ &= \frac{2}{N}\sum_{t=1}^{N}\langle \mathrm{E}(y_t X_t), \widehat{M} - M^\epsilon \rangle - \lambda \langle B^\epsilon, \widehat{M} - M^\epsilon \rangle + \frac{2}{N}\sum_{t=1}^{N}\langle y_t X_t - \mathrm{E}(y_t X_t), \widehat{M} - M^\epsilon \rangle\end{aligned} \tag{C.10}$$

Due to the monotonicity of subdifferentials of convex functions $M \mapsto \|M\|_*$, $\langle \widehat{B} - B^\epsilon, \widehat{M} - M^\epsilon \rangle \geq 0$. So, (C.10) can be further simplified:

$$\begin{aligned}
\frac{2}{np}\langle \widehat{M} - M, \widehat{M} - M^\epsilon \rangle &\leq -\lambda \langle B^\epsilon, \widehat{M} - M^\epsilon \rangle + 2\langle \Delta, \widehat{M} - M^\epsilon \rangle \\
&\underrightarrow{(C.8)} = -\lambda \langle \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T + P_{U^\perp} W P_{V^\perp}, \widehat{M} - M^\epsilon \rangle + 2\langle \Delta, \widehat{M} - M^\epsilon \rangle \\
&\underrightarrow{(C.9)} = -\lambda \langle \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T, \widehat{M} - M^\epsilon \rangle + 2\langle \Delta, \widehat{M} - M^\epsilon \rangle - \lambda \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_*
\end{aligned} \tag{C.11}$$

where $\Delta = N^{-1} \sum_{t=1}^N [y_t X_t - \mathrm{E}(y_t X_t)]$.

By arithmetic, we see that the left-hand side of (C.11) is equal to:

$$\begin{aligned}
2\langle \widehat{M} - M, \widehat{M} - M^\epsilon \rangle &= \langle \widehat{M} - M, \widehat{M} - M + M - M^\epsilon \rangle + \langle \widehat{M} - M^\epsilon + M^\epsilon - M, \widehat{M} - M^\epsilon \rangle \\
&= \|\widehat{M} - M\|_F^2 - \|M^\epsilon - M\|_F^2 + \|\widehat{M} - M^\epsilon\|_F^2.
\end{aligned} \tag{C.12}$$

As for the right side of (C.11), we use the following facts:

$$\| \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T \|_{op} = 1 \quad \text{and} \quad \langle \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T, \widehat{M} - M^\epsilon \rangle = \langle \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T, P_U(\widehat{M} - M^\epsilon) P_V \rangle. \tag{C.13}$$

Given (C.12)-(C.13), (C.11) becomes

$$\begin{aligned}
&\frac{1}{np}\|\widehat{M} - M\|_F^2 + \frac{1}{np}\|\widehat{M} - M^\epsilon\|_F^2 + \lambda \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_* \\
&\leq -\lambda \langle \sum_{j=1}^{J^*(\epsilon)} u_j v_j^T, \widehat{M} - M^\epsilon \rangle + \frac{1}{np}\|M^\epsilon - M\|_F^2 + 2\langle \Delta, \widehat{M} - M^\epsilon \rangle \\
&\leq \lambda \|P_U(M^\epsilon - \widehat{M}) P_V\|_* + \frac{1}{np}\|M^\epsilon - M\|_F^2 + 2\langle \Delta, \widehat{M} - M^\epsilon \rangle,
\end{aligned} \tag{C.14}$$

where the last inequality is due to $|\langle M_1, M_2 \rangle| \leq \|M_1\|_{op} \times \|M_2\|_*$.

In (C.14), the stochastic error term $\langle \Delta, \widehat{M} - M^\epsilon \rangle$ can be decomposed:

$$\begin{aligned}
\langle \Delta, \widehat{M} - M^\epsilon \rangle &= \langle \mathcal{P}_{M^\epsilon}(\Delta), \widehat{M} - M^\epsilon \rangle + \langle P_{U^\perp} \Delta P_{V^\perp}, \widehat{M} - M^\epsilon \rangle \\
&= \langle \mathcal{P}_{M^\epsilon}(\Delta), \mathcal{P}_{M^\epsilon}(\widehat{M} - M^\epsilon) \rangle + \langle \mathcal{P}_{M^\epsilon}(\Delta), \mathcal{P}_{U^\perp}(\widehat{M} - M^\epsilon) P_{V^\perp} \rangle \\
&\quad + \langle P_{U^\perp} \Delta P_{V^\perp}, \mathcal{P}_{M^\epsilon}(\widehat{M}) \rangle + \langle P_{U^\perp} \Delta P_{V^\perp}, P_{U^\perp} \widehat{M} P_{V^\perp} \rangle - \langle P_{U^\perp} \Delta P_{V^\perp}, M^\epsilon \rangle \\
&= \langle \mathcal{P}_{M^\epsilon}(\Delta), \mathcal{P}_{M^\epsilon}(\widehat{M} - M^\epsilon) \rangle + \langle P_{U^\perp} \Delta P_{V^\perp}, P_{U^\perp} \widehat{M} P_{V^\perp} \rangle
\end{aligned} \tag{C.15}$$

where $\mathcal{P}_{M^\epsilon}(\Delta) = \Delta - P_{U^\perp} \Delta P_{V^\perp}$. So it can be upper bounded by:

$$\begin{aligned}
|\langle \Delta, \widehat{M} - M^\epsilon \rangle| &\le \|\mathcal{P}_{M^\epsilon}(\Delta)\|_F \|\mathcal{P}_{M^\epsilon}(\widehat{M} - M^\epsilon)\|_F + \|P_{U^\perp} \Delta P_{V^\perp}\|_{op} \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_* \\
&\le \|\mathcal{P}_{M^\epsilon}(\Delta)\|_F \|\widehat{M} - M^\epsilon\|_F + \|P_{U^\perp} \Delta P_{V^\perp}\|_{op} \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_* \\
&\le \sqrt{2J^*(\epsilon)} \|\Delta\|_{op} \|\widehat{M} - M^\epsilon\|_F + \|\Delta\|_{op} \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_*.
\end{aligned} \tag{C.16}$$

The last inequality is due to the facts that

$$\begin{aligned}
\|\mathcal{P}_{M^\epsilon}(\Delta)\|_F &\le \sqrt{\operatorname{rank}(\mathcal{P}_{M^\epsilon}(\Delta))} \|\Delta\|_{op} = \sqrt{\operatorname{rank}(P_{U^\perp} \Delta P_V + P_U \Delta)} \|\Delta\|_{op} \\
&\le \sqrt{2 \operatorname{rank}(M^\epsilon)} \|\Delta\|_{op} = \sqrt{2J^*(\epsilon)} \|\Delta\|_{op}
\end{aligned}$$

and $\|P_{U^\perp} \Delta P_{V^\perp}\|_{op} \le \|\Delta\|_{op}$.

Meanwhile, the first term in the right-hand side of (C.14) can also be bounded:

$$\|P_U(M^\epsilon - \widehat{M}) P_V\|_* \le \sqrt{\operatorname{rank}(M^\epsilon)} \|P_U(M^\epsilon - \widehat{M}) P_V\|_F \le \sqrt{J^*(\epsilon)} \|M^\epsilon - \widehat{M}\|_F. \tag{C.17}$$

Combining (C.16) - (C.17), (C.14) becomes

$$\begin{aligned}
&\frac{1}{np} \|\widehat{M} - M\|_F^2 + \frac{1}{np} \|\widehat{M} - M^\epsilon\|_F^2 + (\lambda - 2\|\Delta\|_{op}) \|P_{U^\perp} \widehat{M} P_{V^\perp}\|_* \\
&\qquad \le \lambda \sqrt{J^*(\epsilon)} \|M^\epsilon - \widehat{M}\|_F + \epsilon^2 + 2\sqrt{2J^*(\epsilon)} \|\Delta\|_{op} \|\widehat{M} - M^\epsilon\|_F.
\end{aligned} \tag{C.18}$$

If $\lambda \geq 2\|\Delta\|_{op}$, then

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 + \frac{1}{np}\|\widehat{M} - M^\epsilon\|_F^2 \leq \epsilon^2 + (1+\sqrt{2})\lambda\sqrt{J^*(\epsilon)}\|\widehat{M} - M^\epsilon\|_F \tag{C.19}$$

which implies

$$\begin{aligned}\frac{1}{np}\|\widehat{M} - M\|_F^2 &\leq \epsilon^2 + (1+\sqrt{2})\lambda\sqrt{J^*(\epsilon)}\|\widehat{M} - M^\epsilon\|_F - \frac{1}{np}\|\widehat{M} - M^\epsilon\|_F^2 \\ &\leq \epsilon^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 J^*(\epsilon)\lambda^2 np\end{aligned} \tag{C.20}$$

as claimed. □

The result in Lemma 3.5 still contains regularization parameter λ. When λ is selected too large, then entries of $\widehat{M}$ will be overly shrunk toward zero and give poor reconstruction error. If λ is too small, then our constraint, $\lambda \geq 2\|\Delta\|_{op}$, will not be satisfied. Thus, it is important to identify a minimal value for λ such that $\lambda \geq 2\|\Delta\|_{op}$ with high probability. Here, we introduce the second lemma, which gives an upper bound for $\|\Delta\|_{op}$.

Lemma 3.6. *Consider the same data generating mechanism as in Lemma 3.5, with X_t are i.i.d uniformly sampled from $\mathcal{X}$. Then, there exists constant c_1 (dependent on σ and $\|M\|_\infty$) such that*

$$\|\Delta\|_{op} \leq c_1\left[\sqrt{\frac{\log(n+p)}{N(n\wedge p)}} + \sqrt{\log\left(\frac{8(n\wedge p)}{3\sigma^2}\right)\frac{\log(n+p)}{N}}\right] \tag{C.21}$$

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

Furthermore, when $N \geq (n\wedge p)\log^2(n+p)$, we have $\|\Delta\|_{op} \leq 2c_1\sqrt{\frac{\log(n+p)}{N(n\wedge p)}}$ with probability at least $1 - 2(n+p)^{-1}$.

To derive the bound of the stochastic error Δ, we shall use the matrix version of Bernstein's inequality. We now use 2 propositions from [110]. For completeness we include statements of the propositions here below.

Proposition 3.4. *Let* $\{Z_t\}_{t=1}^N$ *be i.i.d* $n \times p$ *matrices that satisfy for some* $\alpha \geq 1$ *and all* t

$$\mathrm{E}\, Z_t = \mathbf{0}, \quad K := \| \|Z_t\|_{op} \|_{\Psi(\alpha)} < \infty,$$

where $\|\cdot\|_{\Psi(\alpha)}$ *is the* $\Psi(\alpha)$*-Orlicz norm defined as* $\|z\|_{\Psi(\alpha)} := \inf \left\{ c > 0 : \mathrm{E} \exp\left(\frac{|z^\alpha|}{c^\alpha}\right) \leq 2 \right\}$ *for a random variable* $z \in \mathbb{R}$*. Define*

$$R^2 := \max \left\{ \left\| \frac{1}{N} \sum_{t=1}^N \mathrm{E}\, Z_t Z_t^T \right\|_{op}, \left\| \frac{1}{N} \sum_{t=1}^N \mathrm{E}\, Z_t^T Z_t \right\|_{op} \right\}.$$

Then for a constant $\widetilde{c}$ *and for all* $h > 0$*,*

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_{t=1}^N Z_t \right\|_{op} \geq \widetilde{c} R \sqrt{\frac{h + \log(n+p)}{N}} + \widetilde{c} \log^{1/\alpha}\left(\frac{K}{R}\right) \left(\frac{h + \log(n+p)}{N} \right) \right) \leq \exp(-h).$$

Proposition 3.5. *Let* $\{Z_t\}_{t=1}^N$ *be* $n \times p$ *matrices that satisfy for a constant* K_1

$$\mathrm{E}\, Z_t = \mathbf{0}, \quad \max_{1 \leq t \leq N} \|Z_t\|_{op} \leq K_1.$$

With the same definition for R *as in Proposition 3.4 Then for all* $h > 0$*,*

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_{t=1}^N Z_t \right\|_{op} \geq \sqrt{2} R \sqrt{\frac{h + \log(n+p)}{N}} + \frac{K_1 [h + \log(n+p)]}{3N} \right) \leq \exp(-h).$$

Given the above results, we now prove Lemma 3.6.

Proof.[Proof of Lemma 3.6]

$\|\Delta\|_{op}$ can be decomposed into two parts as below and we shall bound each part respectively.

$$
\begin{aligned}
\|\Delta\|_{op} &= \left\| \frac{1}{N} \sum_{t=1}^{N} [y_t X_t - \mathrm{E}(y_t X_t)] \right\|_{op} \\
&= \left\| \frac{1}{N} \sum_{t=1}^{N} [\xi_t X_t - \mathrm{E}(\xi_t X_t) + \mathrm{tr}[M^T X_t] X_t - \mathrm{E}(\mathrm{tr}(M^T X_t) X_t) \right\|_{op} \\
&\leq \left\| \frac{1}{N} \sum_{t=1}^{N} \xi_t X_t \right\|_{op} + \left\| \frac{1}{N} \sum_{t=1}^{N} \left(\mathrm{tr}(M^T X_t) X_t - \mathrm{E}(\mathrm{tr}(M^T X_t) X_t) \right) \right\|_{op} \\
&= I_1 + I_2.
\end{aligned}
\tag{C.22}
$$

We use Proposition 3.4 to bound I_1. Let $Z_{1,t} = \xi_t X_t$. Since $\xi_t \sim_{i.i.d} N(0, \sigma^2)$ and X_t are i.i.d uniformly sampled from $\mathcal{X}$ with $\xi_t \perp\!\!\!\perp X_t$, $\{Z_{1,t}\}_{t=1}^N$ are $i.i.d$ $n \times p$ matrices having

$$
\mathrm{E}\, Z_{1,t} = \mathbf{0}, \quad K := \| \|Z_{1,t}\|_{op} \|_{\Psi(\alpha)} = \|\xi_t\|_{\Psi(\alpha)}.
$$

For a normal variable $z \sim N(0,1)$, we have $\mathrm{E} \exp(z^2/c^2) = c/\sqrt{c^2 - 2}$ when $c > \sqrt{2}$. Thus, $\mathrm{E} \exp(z^2/c^2) \leq 2 \Rightarrow c \geq \sqrt{8/3}$. So, $K = \|\xi_t\|_{\Psi(2)} = \sqrt{8/3}$. Let

$$
\begin{aligned}
R^2 &:= \max \left\{ \left\| \frac{1}{N} \sum_{t=1}^{N} \mathrm{E}\, Z_{1,t} Z_{1,t}^T \right\|_{op}, \left\| \frac{1}{N} \sum_{t=1}^{N} \mathrm{E}\, Z_{1,t}^T Z_{1,t} \right\|_{op} \right\} \\
&= \sigma^2 \max \left\{ \left\| \frac{1}{N} \sum_{t=1}^{N} \mathrm{E}(X_t X_t^T) \right\|_{op}, \left\| \frac{1}{N} \sum_{t=1}^{N} \mathrm{E}(X_t^T X_t) \right\|_{op} \right\} \\
&= \frac{\sigma^2}{n \wedge p}.
\end{aligned}
$$

Due to Proposition 3.4, for some $\widetilde{c}$ and for all $h > 0$, we have

$$
\mathbb{P}\left(I_1 \geq \widetilde{c}\sigma \sqrt{\frac{h + \log(n+p)}{N(n \wedge p)}} + \widetilde{c} \sqrt{\frac{1}{2} \log\left(\frac{8(n \wedge p)}{3\sigma^2} \right)} \left(\frac{h + \log(n+p)}{N} \right) \right) \leq \exp(-h).
\tag{C.23}
$$

Similarly, we use Proposition 3.5 to bound I_2. Let $Z_{2,t} = \mathrm{tr}(M^T X_t) X_t - \mathrm{E}(\mathrm{tr}(M^T X_t) X_t)$,

where $\mathrm{E}(\mathrm{tr}(M^T X_t)X_t) = \frac{1}{np}M$. So, $\mathrm{E}(Z_{2,t}) = \mathbf{0}$, and

$$\|Z_{2,t}\|_{op} \leq \left\|\mathrm{tr}(M^T X_t)X_t\right\|_{op} + \left\|\mathrm{E}(\mathrm{tr}(M^T X_t)X_t)\right\|_{op} \leq 2\|M\|_\infty.$$

So, let $K_1 = 2\|M\|_\infty$. Then $\max_{1\leq t\leq N}\|Z_{2,t}\|_{op} \leq K_1$. Consider,

$$\mathrm{E}(Z_{2,t}Z_{2,t}^T) = \mathrm{E}\left[\mathrm{tr}(M^T X)^2 XX^T\right] - \left(\frac{1}{np}\right)^2 MM^T,$$

$$\mathrm{E}(Z_{2,t}^T Z_{2,t}) = \mathrm{E}\left[\mathrm{tr}(M^T X)^2 X^T X\right] - \left(\frac{1}{np}\right)^2 M^T M.$$

Then,

$$\begin{aligned}\left\|\mathrm{E}(Z_{2,t}Z_{2,t}^T)\right\|_{op} &\leq \left\|\mathrm{E}\left[\mathrm{tr}(M^T X)^2 XX^T\right]\right\|_{op} + \left\|\left(\frac{1}{np}\right)^2 MM^T\right\|_{op} \\ &\leq \|M\|_\infty^2/n + \frac{\|M\|_\infty^2}{np} \leq 2\|M\|_\infty^2/n,\end{aligned}$$

and similarly $\left\|\mathrm{E}(Z_{2,t}^T Z_{2,t})\right\|_{op} \leq 2\|M\|_\infty^2/p$. Let

$$R_1^2 := \max\left\{\left\|\frac{1}{N}\sum_{t=1}^N \mathrm{E}(Z_{2,t}Z_{2,t}^T)\right\|_{op}, \left\|\frac{1}{N}\sum_{t=1}^N \mathrm{E}(Z_{2,t}^T Z_{2,t})\right\|_{op}\right\} \leq \frac{2\|M\|_\infty^2}{n\wedge p}.$$

Then, applying Proposition 3.5, we have

$$\mathbb{P}\left(I_2 \geq 2\|M\|_\infty\sqrt{\frac{h+\log(n+p)}{N(n\wedge p)}} + \frac{2\|M\|_\infty[h+\log(n+p)]}{3N}\right) \leq \exp(-h). \tag{C.24}$$

Combining the results of (C.23) and (C.24), for all $h > 0$

$$\begin{aligned}&\mathbb{P}\left(\|\Delta\|_{op} \geq (\widetilde{c}\sigma + 2\|M\|_\infty)\left[\sqrt{\frac{h+\log(n+p)}{N(n\wedge p)}} + \sqrt{\frac{1}{2}\log\left(\frac{8(n\wedge p)}{3\sigma^2}\right)}\left(\frac{h+\log(n+p)}{N}\right)\right]\right) \\ &\leq 2\exp(-h).\end{aligned} \tag{C.25}$$

Select $h = \log(n+p)$ and let $c_1 = \sqrt{2}(\widetilde{c}\sigma + 2\|M\|_\infty)$, then

$$\mathbb{P}\left(\|\Delta\|_{op} \geq c_1\left[\sqrt{\frac{\log(n+p)}{N(n\wedge p)}} + \sqrt{\log\left(\frac{8(n\wedge p)}{3\sigma^2}\right)\frac{\log(n+p)}{N}}\right]\right) \leq 2(n+p)^{-1}. \tag{C.26}$$

In particular, if $N \geq (n\wedge p)\log^2(n+p)$, we have

$$\mathbb{P}\left(\|\Delta\|_{op} \geq 2c_1\sqrt{\frac{\log(n+p)}{N(n\wedge p)}}\right) \leq 2(n+p)^{-1},$$

as desired.

□

Based on Lemma 3.5 and 3.6, it is straightforward to prove Theorem 4.8.

Proof.[Proof of Theorem 4.8]

When $N \geq (n\wedge p)\log^2(n+p)$, we choose λ of the following form

$$\lambda = C_2\sqrt{\frac{\log(n+p)}{N(n\wedge p)}} \tag{C.27}$$

where $C_2 > 0$ is a constant with $C_2 \geq 4c_1$, where c_1 is defined in Lemma 3.6 that only depends on σ and $\|M\|_\infty$. Following from (C.4), then

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 \leq C_2^2\left(\frac{1+\sqrt{2}}{2}\right)^2\frac{(n\vee p)\log(n+p)}{N}J^*(\epsilon) + \epsilon^2. \tag{C.28}$$

holds with probability $1 - 2(n+p)^{-1}$.

This completes the proof. □

C.3 Proof of Lemma 4.4

We begin with an outline of the proof. To form our set of basis functions, we will tessellate our domain $\mathbb{X}^K$ with ∞-norm balls, and use a Taylor series centered at an arbitrary point within each ball to get a uniform approximation for functions in that ball. For a fixed center

point, the Taylor series is a linear combination of fixed basis functions. To obtain our full set of basis functions, we will collect all of the terms in all of those Taylor series. We now formalize this:

Proof. For functions satisfying Condition (4.2), we consider a Taylor series approximation to $f \in \mathcal{F}(L, \gamma, K)$ of order L at a point $\mathbf{x^0} \in \mathbb{R}^K_{[0,1]}$, that is

$$T_{\mathbf{x}^0} f(\mathbf{x}) = f(\mathbf{x^0}) + \sum_{l \leq L-1} \frac{1}{l!} \nabla^l f(\mathbf{x^0})(\mathbf{x} - \mathbf{x^0})^l,$$

where $l! = l_1! \dots l_k!$, $\nabla^l f(\mathbf{x}) = \frac{\partial^l f}{\partial x_1^{l_1} \cdots \partial x_k^{l_k}}$ and $\mathbf{x}^l = x_1^{l_1} \cdots x_k^{l_k}$ over all combinations with $l_1 + \cdots + l_k = l$. There exists $\mathbf{x}' = (x'_1, ..., x'_K)^T \in \mathbb{R}^K_{[0,1]}$ in a neighborhood of radius $\|\mathbf{x} - \mathbf{x}^0\|_2$ centered at $\mathbf{x}^0$ such that the approximation error obeys

$$\begin{aligned}
|f(\mathbf{x}) - T_{\mathbf{x}^0} f(\mathbf{x})| &\leq \left| \sum_{L_1 + \cdots + L_K = L} \frac{1}{L_1! \dots L_k!} \times \frac{\partial^L f(\mathbf{x}')}{\partial {x'_1}^{L_1} \cdots \partial {x'_K}^{L_K}} |x_1 - x_1^0|^{L_1} \cdots |x_K - x_K^0|^{L_K} \right| \\
&\xrightarrow{\text{Condition 4.2}} \leq \gamma \left| \sum_{L_1 + \cdots + L_K = L} \frac{|x_1 - x_1^0|^{L_1} \cdots |x_K - x_K^0|^{L_K}}{L_1! \dots L_k!} \right| \\
&\xrightarrow{\text{Multinomial Theorem}} = \frac{\gamma}{L!} \left(|x_1 - x_1^0| + ... + |x_K - x_K^0| \right)^L
\end{aligned} \tag{C.29}$$

If we consider the approximation error within an ∞-norm ball of radius d (and choose any point in that ball as $\mathbf{x}^0$), then $|x_k - x_k^0| \leq d$ for $k = 1, ..., K$. (C.29) has

$$|f(\mathbf{x}) - T_{\mathbf{x}^0} f(\mathbf{x})| \leq \frac{\gamma}{L!} K^L d^L \tag{C.30}$$

Thus, to get an approximation error of ϵ, let $\frac{\gamma}{L!} K^L d^L = \epsilon$, we need to divide the space into balls of radius

$$d = \sqrt[L]{\frac{L!}{\gamma K^L}} \times \epsilon^{1/L}. \tag{C.31}$$

As the support $\mathbb{R}^K_{[0,1]}$ is bounded by 1, we need $(1/d)^K$ balls with radius d (in ∞-norm) to cover the entirety of $\mathbb{X}^K$, resulting in $\binom{K+L}{L}(1/d)^K$ total terms to get an approximation error

ϵ (above Taylor series approximation contains $\binom{K+L}{L}$ terms). If we select balls of radius d in (C.31), this gives us a total number of terms in our linear expansion

$$J^*(\epsilon) = \binom{K+L}{L} \left(\frac{L!}{\gamma K^L}\right)^{-K/L} \epsilon^{-K/L}.$$

That is, $J^*(\epsilon) = O\left(\epsilon^{-K/L}\right)$. □

C.4 Proof of Theorem 4.9

The proof of this theorem is quite straightforward by connecting a few pieces we have already built.

Proof. Given Condition 4.2 and Lemma 4.4, we have $J^*(\epsilon) = C_3\epsilon^{-\frac{K}{L}}$ for some constant C_3 relying on γ, K, and L. Plugging in this to the upper bound in Theorem 4.8, the upper bound (C.28) then becomes

$$C_2^2 C_3 \left(\frac{1+\sqrt{2}}{2}\right)^2 \frac{(n \vee p)\log(n+p)}{N} \epsilon^{-\frac{K}{L}} + \epsilon^2, \tag{C.32}$$

which is optimized at

$$\begin{aligned} &\frac{(n \vee p)\log(n+p)}{N} \epsilon^{-\frac{K}{L}} = \epsilon^2 \\ \Rightarrow \quad &\epsilon = \left(\frac{(n \vee p)\log(n+p)}{N}\right)^{\frac{L}{2L+K}}. \end{aligned} \tag{C.33}$$

So, we have

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 \leq C^* \left(\frac{(n \vee p)\log(n+p)}{N}\right)^{\frac{2L}{2L+K}} \tag{C.34}$$

with probability at least $1 - 2(n+p)^{-1}$ with $C^* = C_2^2 C_3 \left(\frac{1+\sqrt{2}}{2}\right)^2 + 1$. Equivalently, we can say

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 = O_P\left(\left[\frac{(n \vee p)\log(n+p)}{N}\right]^{\frac{2L}{2L+K}}\right).$$

as claimed. □

C.5 Deriving the Minimax Lower Bound

In this section, we derive the minimax lower bound for estimation within $M(L, \gamma, K)$: We show that the convergence rate in Theorem 4.9 is optimal (up to log terms).

Recall that we assume the true M belongs to the following class of matrices:

$$\mathcal{M}(L, \gamma, K) := \{M \in \mathbb{R}^{n \times p} : m_{ij} = f_j(\boldsymbol{\theta}_{i,\cdot}),\ \boldsymbol{\theta}_{i,\cdot} \in \mathbb{R}^K_{[0,1]}, f_j \in \mathcal{F}(L, \gamma, K), \forall j \leq p\}, \quad \text{(C.35)}$$

where $\mathcal{F}(L, \gamma, K)$ is a class of functions with bounded derivatives:

$$\mathcal{F}(L, \gamma, K) := \left\{ f : \left| \frac{\partial^L}{\partial x_1^{L_1} \cdots x_K^{L_K}} f(\mathbf{x}) \right|_{\mathbf{x}=\mathbf{x^0}} \right| \leq \gamma,\ \forall \mathbf{x}^0 \in \mathbb{R}^K_{[0,1]},\ \sum_{k=1}^K L_k = L \right\}. \quad \text{(C.36)}$$

For simplicity of notation, let $\boldsymbol{\theta}_i := \boldsymbol{\theta}_{i,\cdot} \in \mathbb{R}^K_{[0,1]}$ denote the i-th row vector of the embeddings $\Theta \in \mathbb{R}^{n \times K}$ in this section.

We shall obtain the lower bound based on information theory. The bound is with respect to $\|\cdot\|_F^2$-risk. We pose things in terms of the error in a multi-way hypothesis testing problem, where the set of testing hypotheses should be a suitably large packing set for $\mathcal{M}(L, \gamma, K)$. In this section, we first show the existence of such a suitably large packing set. Then, we apply Yang's method [127] to prove the main results in Theorem 4.10.

C.5.1 Constructing the $2\delta_{N,n,p}$-packing Set

For $M \in \mathcal{M}(L, \gamma, K)$, the risk of the estimator can be written as

$$\frac{1}{np}\|\widehat{M} - M\|_F^2 = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \left[\widehat{m}_{ij} - f_j(\boldsymbol{\theta}_i)\right]^2.$$

This is to say, bounding $\frac{1}{np}\|\widehat{M} - M\|_F^2$ can be viewed as a classical nonparametric regression problem. So, we follow the construction of many hypotheses in Section 2.6 of [107]. However, here we are working in a multi-dimensional setting, i.e. $\boldsymbol{\theta}_i \in \mathbb{R}^K$, $K \geq 1$.

In giving our packing set, we will work with combinations of "bump functions". To define these, we need an archetypal ingredient — the bump functions that we will use:

$$\varphi(u) = c_L e \times \exp\left(-\frac{1}{1-4u^2}\right), \quad u \in (-1/2, 1/2), \tag{C.37}$$

which is infinitely differentiable and vanishes outside of $(-1/2, 1/2)$. $c_L > 0$ is a tiny constant that only depends on L such that $|\partial^l \varphi(u)/\partial u^l| \leq 1$, $\forall l = 0, 1, ..., L$. Meanwhile, since $\int_{-1/2}^{1/2} e^2 \exp^2\left(-\frac{1}{1-4u^2}\right) du > 0.49$, (it is actually very close to 0.5), we have $\|\varphi\|_2^2 := \int_{-1/2}^{1/2} \varphi^2(u) du > 0.49 c_L^2$. In addition, the maximum value of this function is $\sup_u |\varphi(u)| = \varphi(0) = c_L$.

Now, we shall work under the multidimensional setting. We use bold letters to refer to multivariate indices and regular letters to refer to the indices of each coordinate. Let $\boldsymbol{i} = (i_1, ..., i_K) \in \{1, 2, ..., \sqrt[K]{n}\}^K$ having $\sum_{\boldsymbol{i}=(i_1,...,i_K)} 1 = \sum_{i_1=1}^{\sqrt[K]{n}} ... \sum_{i_K=1}^{\sqrt[K]{n}} 1 = n$, where $\sqrt[K]{n}$ is assumed to be an integer. Suppose that the observed embeddings follows a fixed equispaced design, i.e. $\boldsymbol{\theta_i} = \boldsymbol{\theta}_{(i_1,...,i_K)} = (\theta_{i_1}, ..., \theta_{i_K})^T = (\frac{i_1}{\sqrt[K]{n}}, ..., \frac{i_K}{\sqrt[K]{n}})^T$. Consider a multivariate function $\Phi_{\boldsymbol{d}} : \mathbb{R}^K \to \mathbb{R}$,

$$\begin{aligned} \Phi_{\boldsymbol{d}}(\boldsymbol{\theta_i}) &= \gamma b^{-L/K} \prod_{k=1}^{K} \varphi_{d_k}(\theta_{i_k}) \\ &:= \gamma b^{-L/K} \prod_{k=1}^{K} \varphi(\sqrt[K]{b}\theta_{i_k} - d_k + 1/2), \end{aligned} \tag{C.38}$$

where $\boldsymbol{d} = (d_1, ..., d_K) \in \{1, 2, ..., \sqrt[K]{b}\}^K$. Here $b \geq 1$ is an integer that depends on N, n, p and some constant c_0, and will be specified later. $\varphi(u)$ is defined in (C.37). Then, we have the following technical lemma for $\Phi_{\boldsymbol{d}}$, which will later be used for constructing the packing set.

Lemma 3.7. *Suppose $\varphi(\cdot)$ are given by (C.37). Then, $\Phi_{\boldsymbol{d}}$ has the following properties:*

(i) $\Phi_{\boldsymbol{d}}(\mathbf{x}) \in \mathcal{F}(L, \gamma, K)$.

(ii) $\Phi_{\boldsymbol{d}}$ *have disjoint support for different* $\boldsymbol{d}$.

(iii) There exist $C_{1,L,K} > 0$ *and* $C_{2,L,K} > 0$ *only dependent on* L *and* K*, for any given* $\boldsymbol{d}$*,* $\Phi_{\boldsymbol{d}}$ *has*

$$\gamma^2 C_{2,L,K} b^{-\frac{2L+K}{K}} \leq \frac{1}{n} \sum_{\boldsymbol{i}=(i_1,...,i_K)} \Phi^2_{\boldsymbol{d}}(\boldsymbol{\theta_i}) \leq \gamma^2 C_{1,L,K} b^{-\frac{2L+K}{K}}$$

when integer b *satisfies* $1 \leq b \leq 0.48^K n$*.*

Proof. For $\varphi(\cdot)$ in (C.37), we have $|\frac{\partial^l}{\partial u^l}\varphi(u)| \leq 1$, $\forall l = 0, 1, ..., L$ such that $\left|\frac{\partial^L}{\partial x_1^{L_1} ... x_K^{L_K}} \Phi_{\boldsymbol{d}}(\mathbf{x})\right| \leq \gamma$ holds for any $L_1 + ... + L_K = L$ for $\mathbf{x} \in \mathbb{R}^K_{[0,1]}$. Thus, $\Phi_{\boldsymbol{d}}(\mathbf{x}) \in \mathcal{F}(L, \gamma, K)$.

Given that $\varphi(u) > 0$ if and only if $u \in (-1/2, 1/2)$, we have $\varphi_{d_k}(x) \equiv \varphi(\sqrt[K]{b}x - d_k + 1/2) > 0$ if and only if $x \in \left(\frac{d_k - 1}{\sqrt[K]{b}}, \frac{d_k}{\sqrt[K]{b}}\right)$ for $d_k \in \{1, ..., \sqrt[K]{b}\}$. So, for each, we can divide the space $[0, 1]$ into $\sqrt[K]{b}$ intervals, i.e.

$$\Delta_1 = \left[0, \frac{1}{\sqrt[K]{b}}\right], \quad \Delta_{d_k} = \left(\frac{d_k - 1}{\sqrt[K]{b}}, \frac{d_k}{\sqrt[K]{b}}\right], \ d_k = 2, ..., \sqrt[K]{b},$$

such that $\Delta_{d_k} \cap \Delta_{d'_k} = \emptyset$ for $d_k \neq d'_k$ and $\cup_{d_k} \Delta_{d_k} = [0, 1]$. Thus, $\varphi_{d_k}(x)$ have disjoint support and their support union is the unit interval.

Because $\Phi_{\boldsymbol{d}}$ is the product of φ_{d_k}, they also have disjoint supports. That is, for each $\boldsymbol{d}$, $\Phi_{\boldsymbol{d}}(\mathbf{x}) > 0$ only when $\mathbf{x} \in \Delta_{\boldsymbol{d}}$ where

$$\begin{aligned}
\Delta_{\boldsymbol{d}=(1,1,...,1)} &= \left[0, \frac{1}{\sqrt[K]{b}}\right] \times ... \times \left[0, \frac{1}{\sqrt[K]{b}}\right], \\
\Delta_{\boldsymbol{d}=(d_1,...d_K)} &= \left(\frac{d_1 - 1}{\sqrt[K]{b}}, \frac{d_1}{\sqrt[K]{b}}\right] \times ... \times \left(\frac{d_K - 1}{\sqrt[K]{b}}, \frac{d_K}{\sqrt[K]{b}}\right],
\end{aligned}$$

$d_k = 2, ..., \sqrt[K]{b}$ for $k = 1, ..., K$, such that $\Delta_{\boldsymbol{d}} \cap \Delta_{\boldsymbol{d}'} = \emptyset$ if $\boldsymbol{d} \neq \boldsymbol{d}'$ and $\cup_{\boldsymbol{d}} \Delta_{\boldsymbol{d}} = [0, 1]^K$. So, the space $[0, 1]^K$ is divided into b disjoint cubes.

As for *(iii)*, we know there exists a constant c_L that only depends on L such that

$\sup_u |\varphi(u)| = \varphi(0) = c_L$, and $\|\varphi\|_2^2 > 0.49 c_L^2$. Then

$$
\begin{aligned}
\frac{1}{n}\sum_{\boldsymbol{i}=(i_1,\ldots,i_K)} \Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i}) &= \frac{1}{n}\gamma^2 b^{-2L/K} \sum_{i_1=1}^{\sqrt[K]{b}} \cdots \sum_{i_K=1}^{\sqrt[K]{b}} \prod_{k=1}^{K} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_k - d_k + 1/2 \right) \\
&= \frac{1}{n}\gamma^2 b^{-2L/K} \sum_{i_2=1}^{\sqrt[K]{b}} \cdots \sum_{i_K=1}^{\sqrt[K]{b}} \prod_{k=2}^{K} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_k - d_k + 1/2 \right) \\
&\qquad \times \left[\sum_{i_1=1}^{\sqrt[K]{b}} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_1 - d_1 + 1/2 \right) \right] \\
&= \frac{1}{n}\gamma^2 b^{-2L/K} \sum_{i_3=1}^{\sqrt[K]{b}} \cdots \sum_{i_K=1}^{\sqrt[K]{b}} \prod_{k=3}^{K} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_k - d_k + 1/2 \right) \\
&\qquad \times \left[\sum_{i_1=1}^{\sqrt[K]{b}} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_1 - d_1 + 1/2 \right) \right] \times \left[\sum_{i_2=1}^{\sqrt[K]{b}} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_2 - d_2 + 1/2 \right) \right] \\
&\cdots \\
&= \frac{1}{n}\gamma^2 b^{-2L/K} \prod_{k=1}^{K} \left\{ \sum_{i_k=1}^{\sqrt[K]{n}} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_k - d_k + 1/2 \right) \right\} \\
&= \frac{1}{n}\gamma^2 b^{-2L/K} \prod_{k=1}^{K} \left\{ \sum_{\sqrt[K]{\frac{n}{b}}(d_k-1) < i_k \le \sqrt[K]{\frac{n}{b}} d_k} \varphi^2\left(\sqrt[K]{\frac{b}{n}} i_k - d_k + 1/2 \right) \right\} \\
&\le \frac{1}{n}\gamma^2 b^{-2L/K} \prod_{k=1}^{K} \left\{ \sqrt[K]{\frac{n}{b}} \times \varphi^2(0) \right\} \\
&= \gamma^2 b^{-\frac{2L+K}{K}} c_L^{2K}
\end{aligned}
\tag{C.39}
$$

Therefore, $\frac{1}{n}\sum_{\boldsymbol{i}=(i_1,\ldots,i_K)} \Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i}) \le c_L^{2K}\gamma^2 b^{-\frac{2L+K}{K}}$. c_L^{2K} is the constant we find for $C_{1,L,K}$.

On the other hand, we use the fact that the upper Riemann sum is greater than the

integral of the function. Thus, for each coordinate,

$$\begin{aligned}
\sqrt[K]{\frac{b}{n}}\sum_{i_k=1}^{\sqrt[K]{n}}\varphi^2\left(\sqrt[K]{\frac{b}{n}}i_k - d_k + 1/2\right) &= \sqrt[K]{\frac{b}{n}}\sum_{\sqrt[K]{\frac{n}{b}}(d_k-1)<i_k\le\sqrt[K]{\frac{n}{b}}d_k}\varphi^2\left[\sqrt[K]{\frac{b}{n}}\left(i_k - \frac{\sqrt[K]{n}(d_k-1/2)}{\sqrt[K]{b}}\right)\right] \\
&= \sqrt[K]{\frac{b}{n}}\sum_{-\frac{1}{2}\frac{\sqrt[K]{n}}{\sqrt[K]{b}}<t\le\frac{1}{2}\frac{\sqrt[K]{n}}{\sqrt[K]{b}}}\varphi^2(\sqrt[K]{b/n}\times t) \\
\underrightarrow{[\varphi(-u)=\varphi(u)]} &= \sqrt[K]{\frac{b}{n}}\left[2\sum_{0\le t\le\frac{1}{2}\frac{\sqrt[K]{n}}{\sqrt[K]{b}}}\varphi^2(\sqrt[K]{b/n}\times t) - \varphi^2(0)\right] \\
\underrightarrow{[\varphi(u)\text{ decreases for }u\ge 0]} &\ge 2\sqrt[K]{\frac{b}{n}}\int_0^{\frac{1}{2}\frac{\sqrt[K]{n}}{\sqrt[K]{b}}}\varphi^2(\sqrt[K]{b/n}\times t)dt - \sqrt[K]{\frac{b}{n}}\varphi^2(0) \\
&= 2\int_0^{1/2}\varphi^2(u)du - \sqrt[K]{\frac{b}{n}}c_L^2 \\
&= \|\varphi\|_2^2 - \sqrt[K]{\frac{b}{n}}c_L^2 \\
&> c_L^2\left(0.49 - \sqrt[K]{b/n}\right) \\
\underrightarrow{\left[1\le b\le 0.49^K n\right]} &\ge 0,
\end{aligned}$$

Thus, the empirical sum can also be lower bounded by

$$\begin{aligned}
\frac{1}{n}\sum_{\boldsymbol{i}=(i_1,\ldots,i_K)}\Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i}) &= \gamma^2 b^{-\frac{2L+K}{K}}\prod_{k=1}^{K}\left\{\sqrt[K]{\frac{b}{n}}\sum_{i_k=1}^{\sqrt[K]{n}}\varphi^2\left(\sqrt[K]{\frac{b}{n}}i_k - d_k + 1/2\right)\right\} \\
&\ge \gamma^2 b^{-\frac{2L+K}{K}}c_L^{2K}\left(0.49 - \sqrt[K]{b/n}\right)^K
\end{aligned} \tag{C.40}$$

When $1\le b\le 0.48^K n$, then

$$\frac{1}{n}\sum_{\boldsymbol{i}=(i_1,\ldots,i_K)}\Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i}) \ge \gamma^2(0.1c_L)^{2K}b^{-\frac{2L+K}{K}}, \tag{C.41}$$

and thus $(0.1c_L)^{2K}$ is the constant we find for $C_{2,L,K}/$

Now combining (C.39) and (C.41), we have

$$\gamma^2 C_{2,L,K} b^{-\frac{2L+K}{K}} \leq \frac{1}{n} \sum_{\boldsymbol{i}=(i_1,\ldots,i_K)} \Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i}) \leq \gamma^2 C_{1,L,K} b^{-\frac{2L+K}{K}}$$

as claimed. □

In proving the lower bound, we shall use Fano's method (see Section 15.3.2 in [116]). To do that, we first establish the connection between minimax risks and error probabilities in testing problems (for completeness), and then apply Fano's inequality to lower bound the error probabilities. To this end, we first provide the following lemma, which shows that there exists a packing set of hypotheses with suitably large cardinality, for which the mutual information (stated in terms of Kullback-Leibler divergence) can be upper bounded. We can then Fano's inequality with this set.

Lemma 3.8. *For this result we consider an arbitrary fixed L, γ and K. For some constant $C_{1,L,K}$ and $C_{2,L,K}$ that only depends on L and K, and for some other constant $c_0 > 0$, there exists a subset $\mathcal{B}^0 \subseteq \mathcal{M}(L, \gamma, K)$ with cardinality*

$$|\mathcal{B}^0| \geq 2^{\lceil c_0 \left(\frac{n \vee p}{N}\right)^{\frac{-K}{2L+K}} \rceil \times p/8} + 1,$$

when $p \geq 8$, that has the following properties:

(i) $\mathcal{B}^0$ is a $2\delta_{N,n,p}$-packing set, i.e. for any $M_s \neq M_{s'} \in \mathcal{B}^0$,

$$\frac{1}{np}\|M_s - M_{s'}\|_F^2 \geq 2\delta_{N,n,p} = \frac{C_{2,L,K}\gamma^2}{8}(2c_0)^{-2L/K}\left(\frac{n \vee p}{N}\right)^{\frac{2L}{2L+K}}$$

when $c_0^{-\frac{2L+K}{K}}(n \vee p) \leq N \leq (2c_0)^{-\frac{2L+K}{K}} 0.48^{2L+K}(n \vee p) n^{\frac{2L+K}{K}}$.

(ii) For any $M_s, M_{s'} \in \mathcal{B}^0$,

$$K(\mathbb{P}_s||\mathbb{P}_{s'}) \leq \frac{C_{1,L,K}\gamma^2}{2\sigma^2} c_0^{-\frac{2L}{K}} N \left(\frac{n \vee p}{N}\right)^{\frac{2L}{2L+K}}$$

where $K(\mathbb{P}_s||\mathbb{P}_{s'})$ denotes the Kullback-Leibler divergence between probability distributions of observations $\{(y_t, X_t)\}_{t=1}^N$ satisfying model (C.2), given M_s and $M_{s'}$ respectively.

Proof. We will consider a positive integer b which depends on N, n, p and a constant c_0. The precise specification of b will come later. Consider the multivariate function $\Phi_{\boldsymbol{d}}(\boldsymbol{\theta})$ in (C.38).

We will define a set Ω that is used to construct packing matrices where each element ω in Ω is a sequence (of length b) of diagonal matrices. We index the set in a somewhat curious way: We use a multi-index of dimension K where each index has elements in $\{1, ..., \sqrt[K]{b}\}$. This will ease exposition later.

$$\boldsymbol{\Omega} = \left\{ \boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{d}})_{\boldsymbol{d} \in \{1,..., \sqrt[K]{b}\}^K} : \text{for each } \boldsymbol{d},\ \boldsymbol{w}_{\boldsymbol{d}} = \text{diag}(w_{\boldsymbol{d},1}, ..., w_{\boldsymbol{d},p}), w_{\boldsymbol{d},j} \in \{0,1\} \right\}, \quad \text{(C.42)}$$

From this we define the following collection of matrices,

$$\begin{aligned}
\mathcal{B} &= \left\{ M_{\boldsymbol{w}} = \sum_{d_1=1}^{\sqrt[K]{b}} \cdots \sum_{d_K=1}^{\sqrt[K]{b}} \begin{pmatrix} \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_1) w_{\boldsymbol{d},1} & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_1) w_{\boldsymbol{d},2} & \dots & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_1) w_{\boldsymbol{d},p} \\ \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_2) w_{\boldsymbol{d},1} & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_2) w_{\boldsymbol{d},2} & \dots & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_2) w_{\boldsymbol{d},p} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_n) w_{\boldsymbol{d},1} & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_n) w_{\boldsymbol{d},2} & \dots & \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_n) w_{\boldsymbol{d},p} \end{pmatrix}_{n \times p}, \quad w_{\boldsymbol{d},j} \in \{0,1\} \right\} \\
&=: \left\{ M_{\boldsymbol{w}} = \sum_{d_1,...,d_K} \Phi_{\boldsymbol{d}}(\boldsymbol{\Theta}) \boldsymbol{w}_{\boldsymbol{d}},\ \text{for } \boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{d}}) \in \boldsymbol{\Omega} \right\}.
\end{aligned} \quad \text{(C.43)}$$

We see that we can compactly write each matrix in our set as the product of $\Phi_{\boldsymbol{d}}(\boldsymbol{\Theta})$ and $\boldsymbol{w}_{\boldsymbol{d}}$, where $\Phi_{\boldsymbol{d}}(\boldsymbol{\Theta})$ is a $n \times p$ matrix whose elements in the i-th row are all $\Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_i)$. It is direct to check that the cardinality of Ω is given by $|\boldsymbol{\Omega}| = |\mathcal{B}| = 2^{bp}$.

Thus, entries of $M_{\boldsymbol{w}} \in \mathcal{B}$ can be written as $m_{ij} = \sum_{d_1,...,d_K} \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_i) w_{\boldsymbol{d},j} = g_j(\boldsymbol{\theta}_i)$, where g_j has bounded derivatives,

$$\left| \frac{\partial^L g_j(\mathbf{x})}{\partial x_1^{L_1} \cdots x_K^{L_K}} \right| \leq \sum_{d_1,...,d_K} \left| \frac{\partial^L \Phi_{\boldsymbol{d}}(\mathbf{x})}{\partial x_1^{L_1} \cdots x_K^{L_K}} \right| = \left| \frac{\partial^L \Phi_{\boldsymbol{d}}(\mathbf{x})}{\partial x_1^{L_1} \cdots x_K^{L_K}} \right| \mathbb{1}\{\mathbf{x} \in \Delta_{\boldsymbol{d}}\} \leq \gamma$$

for $\forall \mathbf{x} \in \mathbb{R}^K_{[0,1]}$. Hence, $\mathcal{B} \subseteq \mathcal{M}(L, \gamma, K)$.

Consider a set of testing hypotheses from $\mathcal{B}$,

$$\mathcal{B}^0 = \{M_{\boldsymbol{w}^{(0)}}, ..., M_{\boldsymbol{w}^{(S)}}\} \subseteq \mathcal{B}, \quad \boldsymbol{w}^{(s)} \in \boldsymbol{\Omega}, \ s = 0, 1, ..., S, \tag{C.44}$$

where $\boldsymbol{w}^{(s)} \neq \boldsymbol{w}^{(s')}$ for $0 \leq s \neq s' \leq S$.

For any $0 \leq s \neq s' \leq S$, and constant $C_{2,L,K}$ only dependent on L,

$$\begin{aligned}
\frac{1}{np}\|M_{\boldsymbol{w}^{(s)}} - M_{\boldsymbol{w}^{(s')}}\|_F^2 &= \frac{1}{np} \sum_{\boldsymbol{i}=(i_1,...,i_K)} \sum_{j=1}^{p} \left[\sum_{d_1=1}^{\sqrt[K]{b}} \cdots \sum_{d_K=1}^{\sqrt[K]{b}} (w^{(s)}_{\boldsymbol{d},j} - w^{(s')}_{\boldsymbol{d},j}) \Phi_{\boldsymbol{d}}(\boldsymbol{\theta}_{\boldsymbol{i}}) \right]^2 \\
\xrightarrow{\text{the support of } \Phi_{\boldsymbol{d}}\text{'s are disjoint}} &= \frac{1}{p} \sum_{j=1}^{p} \sum_{d_1=1}^{\sqrt[K]{b}} \cdots \sum_{d_K=1}^{\sqrt[K]{b}} (w^{(s)}_{\boldsymbol{d},j} - w^{(s')}_{\boldsymbol{d},j})^2 \left(\frac{1}{n} \sum_{\boldsymbol{i}=(i_1,...,i_K)} \Phi^2_{\boldsymbol{d}}(\boldsymbol{\theta}_{\boldsymbol{i}}) \right) \\
\xrightarrow{\text{Lemma 3.7-(i)}} &\geq \gamma^2 C_{2,L,K} b^{-\frac{2L+K}{K}} p^{-1} \rho(\boldsymbol{w}^{(s)}, \boldsymbol{w}^{(s')})
\end{aligned} \tag{C.45}$$

where $\rho(\boldsymbol{w}^{(s)}, \boldsymbol{w}^{(s')}) = \sum_{j=1}^{p} \sum_{d_1=1}^{\sqrt[K]{b}} \cdots \sum_{d_K=1}^{\sqrt[K]{b}} (w^{(s)}_{\boldsymbol{d},j} - w^{(s')}_{\boldsymbol{d},j})^2$ is the hamming distance between $\boldsymbol{w}^{(s)}$ and $\boldsymbol{w}^{(s')}$.

Due to the Varshamov–Gilbert bound (Lemma 2.9 in [107]), when $bp \geq 8$, there exists a subset $\boldsymbol{\Omega}^0 = (\boldsymbol{w}^{(0)}, ..., \boldsymbol{w}^{(S)}) \subseteq \boldsymbol{\Omega}$ such that $S \geq 2^{bp/8}$ and $\rho(\boldsymbol{w}^{(s)}, \boldsymbol{w}^{(s')}) \geq bp/8$ for $0 \leq s \neq s' \leq S$. Since $b \geq 1$, $p \geq 8$ is a sufficient condition to guarantee $bp \geq 8$.

Now, in particular, we choose our testing set based on $\boldsymbol{\Omega}^0$: That is, we place $M_{\boldsymbol{w}^{(s)}} \in \mathcal{B}^0$ if and only if $\boldsymbol{w}^{(s)} \in \boldsymbol{\Omega}^0$. In particular this gives us that $\rho(\boldsymbol{w}^{(s)}, \boldsymbol{w}^{(s')}) \geq bp/8$. for all $w^{(s)}, w^{(s')} \in \mathcal{B}^0$ with $s \neq s'$. Then, following (C.45), we have that

$$\frac{1}{np}\|M_{\boldsymbol{w}^{(s)}} - M_{\boldsymbol{w}^{(s')}}\|_F^2 \geq \frac{\gamma^2 C_{2,L,K}}{8} b^{-\frac{2L}{K}}. \tag{C.46}$$

Now, we finally give the value that we use for b: Select $b = \left\lceil c_0 \left(\frac{n \vee p}{N}\right)^{\frac{-K}{2L+K}} \right\rceil$ for some constant $c_0 > 0$. We note that (C.45)-(C.46) hold only when $b \leq 0.48^K n$ as stated in

Lemma 3.7. So, we need

$$N \leq c_0^{-\frac{2L+K}{K}} 0.48^{2L+K} (n \vee p) n^{\frac{2L+K}{K}}. \tag{C.47}$$

Furthermore, we also need

$$N \geq c_0^{-\frac{2L+K}{K}} (n \vee p) \tag{C.48}$$

such that

$$b = \left\lceil c_0 \left(\frac{n \vee p}{N} \right)^{\frac{-K}{2L+K}} \right\rceil \leq 2c_0 \left(\frac{n \vee p}{N} \right)^{\frac{-K}{2L+K}}. \tag{C.49}$$

This, finally gives us

$$\frac{1}{np} \| M_{\boldsymbol{w}^{(s)}} - M_{\boldsymbol{w}^{(s')}} \|_F^2 \geq \frac{C_{2,L,K} \gamma^2}{8} (2c_0)^{-2L/K} \left(\frac{n \vee p}{N} \right)^{\frac{2L}{2L+K}} =: 2\delta_{N,n,p}. \tag{C.50}$$

Then $\mathcal{B}^0$ is a $2\delta_{N,n,p}$-packing set of $\mathcal{M}(L, \gamma, K)$ and the cardinality $|\mathcal{B}^0| = S + 1 \geq 2^{bp/8} + 1 = 2^{\lceil c_0 \left(\frac{n \vee p}{N} \right)^{\frac{-K}{2L+K}} \rceil \times p/8} + 1$ when $p \geq 8$.

We now show the second property (related to the KL distance) of $\mathcal{B}^0$. For any matrices

$M_{\boldsymbol{w}^{(s)}}, M_{\boldsymbol{w}^{(s')}} \in \mathcal{B}^0$, with the selected $b = \lceil c_0 \left(\frac{n \vee p}{N}\right)^{\frac{-K}{2L+K}} \rceil$, we have

$$
\begin{aligned}
K(\mathbb{P}_s || \mathbb{P}_{s'}) &= \int \log \frac{d\mathbb{P}_s}{d\mathbb{P}_{s'}} d\mathbb{P}_s \\
&= \int \int \log \frac{\prod_{t=1}^N p(y_t, X_t | M_{w^{(s)}})}{\prod_{t=1}^N p(y_t, X_t | M_{w^{(s')}})} \left[\prod_{t=1}^N p(y_t, X_t | M_{w^{(s)}}) dy_t dX_t\right] \\
\xrightarrow{\text{Bayes' rule}} &= \mathrm{E}_{X \sim \Pi} \sum_{t=1}^N \int \left[\log p(y_t | X_t, M_{w^{(s)}}) - \log p(y_t | X_t, M_{w^{(s')}})\right] p(y_t | X_t, M_{w^{(s)}}) dy_t \\
\xrightarrow{(y_t | X_t, M) \sim_{i.i.d} N[\langle X_t, M \rangle, \sigma^2]} &= \mathrm{E}_{X \sim \Pi} \sum_{t=1}^N \frac{\langle X_t, M_{w^{(s)}} - M_{w^{(s')}} \rangle^2}{2\sigma^2} \\
\xrightarrow{\left[E_{X \sim \Pi} \langle X_t, M \rangle^2 = \frac{1}{np} \|M\|_F^2\right]} &= \frac{N}{2\sigma^2 np} \|M_{\boldsymbol{w}^{(s)}} - M_{\boldsymbol{w}^{(s')}}\|_F^2 \\
&\leq \frac{N}{2\sigma^2} \sum_{d_1=1}^{\sqrt[K]{b}} \cdots \sum_{d_K=1}^{\sqrt[K]{b}} \left(\frac{1}{n} \sum_{\boldsymbol{i}=(i_1,\ldots,i_K)} \Phi_{\boldsymbol{d}}^2(\boldsymbol{\theta_i})\right) \\
&\leq \frac{N\gamma^2}{2\sigma^2} b^{-\frac{2L}{K}} C_{1,L,K} \\
&\leq \frac{C_{1,L,K}\gamma^2}{2\sigma^2} c_0^{-\frac{2L}{K}} N \left(\frac{n \vee p}{N}\right)^{\frac{2L}{2L+K}}
\end{aligned}
\tag{C.51}
$$

Thus, Lemma 3.8 is proved.

□

C.5.2 Information-theoretic lower bounds

Given Lemma 3.8, we now apply the argument in [127] to yield a lower bound for error in our estimation problem with respect to Frobenius norm.

Proof.[Proof of Theorem 4.10]

For a given $\delta_{N,n,p}$, let $\mathcal{B}^0$ be the $2\delta_{N,n,p}$-packing set of $\mathcal{M}(L, \gamma, K)$ indicated by Lemma 3.8. We know that for any $M_s \neq M_{s'} \in \mathcal{B}^0$,

$$
\frac{1}{np} \|M_s - M_{s'}\|_F^2 \geq 2\delta_{N,n,p}
$$

with $\delta_{N,n,p} = \frac{C_{2,L,K}\gamma^2}{16}(2c_0)^{-2L/K}\left(\frac{n\vee p}{N}\right)^{\frac{2L}{2L+K}}$, when $c_0^{-\frac{2L+K}{K}}(n \vee p) \leq N \leq (2c_0)^{-\frac{2L+K}{K}}0.48^{2L+K}(n\vee p)n^{\frac{2L+K}{K}}$ for some constant $c_0 > 0$.

Let $d(M_1, M_2) = \frac{1}{np}\|M_1 - M_2\|_F^2$ and define

$$\widetilde{M} = \arg\min_{M'\in\mathcal{B}^0} d(M', \widehat{M}) \in \mathcal{B}^0.$$

Let M be any matrix in the packing set $\mathcal{B}^0$. If $d(M, \widehat{M}) < \delta_{N,n,p}$, then $\max\{d(M, \widehat{M}), d(\widetilde{M}, \widehat{M})\} = d(M, \widehat{M}) < \delta_{N,n,p} \leq \delta_0 \equiv C_{2,L,K}\gamma^2 4^{-(L+2K)/K}$. Then, by the triangle inequality, we have $d(M, \widehat{M}) + d(\widetilde{M}, \widehat{M}) \geq d(M, \widetilde{M}) \geq 2\delta_{N,n,p}$ when $M \neq \widetilde{M}$. This implies that $d(M, \widehat{M}) \geq \delta_{N,n,p}$, which contradicts $d(M, \widehat{M}) < \delta_{N,n,p}$. Therefore, if $M \neq \widetilde{M}$, we must have $d(M, \widehat{M}) \geq \delta_{N,n,p}$. So, it follows that

$$\begin{aligned}\inf_{\widehat{M}}\sup_{M\in\mathcal{M}(L,\gamma,K)}\mathbb{P}\left\{d(M,\widehat{M}) \geq \delta_{N,n,p}\right\} &\geq \inf_{\widehat{M}}\sup_{M\in\mathcal{B}^0}\mathbb{P}\left\{d(M,\widehat{M}) \geq \delta_{N,n,p}\right\} \\ &= \inf_{\widehat{M}}\sup_{M\in\mathcal{B}^0}\mathbb{P}\left\{M \neq \widetilde{M}\right\} \\ &\geq \inf_{\widehat{M}}\mathbb{P}(M \neq \widetilde{M})\end{aligned} \tag{C.52}$$

where M is uniformly distributed over the $2\delta_{N,n,p}$-packing set $\mathcal{B}^0$ with $|\mathcal{B}^0| \geq 2^{\lceil c_0\left(\frac{n\vee p}{N}\right)^{\frac{-K}{2L+K}}\rceil\times p/8} + 1$ as in Lemma 3.8. this has reduced our problem essentially to a testing problem.

We now use this to obtain a lower bound, by considering KL-divergence here. By Lemma 3.8 -(iii), Fano's inequality [29] or [116, Proposition 15.12] and the convexity of the

Kullback–Leibler divergence [116, (15.34)],

$$\begin{aligned}\mathbb{P}(M \neq \widetilde{M}) &\geq 1 - \frac{\frac{1}{|\mathcal{B}^0|^2}\sum_{M_s, M_{s'} \in \mathcal{B}^0} K(\mathbb{P}_s || \mathbb{P}_{s'}) + \log 2}{\log |\mathcal{B}^0|} \\ \xrightarrow{\text{Lemma 3.8}} &\geq 1 - \frac{\frac{C_{1,L,K}\gamma^2}{2\sigma^2} c_0^{-\frac{2L}{K}} N \left(\frac{n \vee p}{N}\right)^{\frac{2L}{2L+K}} + \log 2}{\lceil c_0 \left(\frac{n \vee p}{N}\right)^{\frac{-K}{2L+K}} \rceil p \log 2} \\ \xrightarrow{bp \geq 8} &\geq \frac{7}{8} - \frac{C_{1,L,K}\gamma^2 c_0^{-\frac{2L+K}{K}} (n \vee p)}{2(\log 2)\sigma^2 p}.\end{aligned} \tag{C.53}$$

Consider $n = \kappa p$ for some $\kappa > 0$. Let

$$c_0 = \left(\frac{4 \max(\kappa, 1)\gamma^2 C_{1,L,K}}{3 \log 2\sigma^2}\right)^{\frac{K}{2L+K}}. \tag{C.54}$$

Then,

$$\cdot\mathbb{P}(M \neq \widetilde{M}) \geq 7/8 - \frac{\gamma^2 C_{1,L,K} \max(\kappa, 1) c_0^{-\frac{2L+K}{K}}}{2(\log 2)\sigma^2} = 7/8 - 3/8 = 1/2 \tag{C.55}$$

Thus, it follows from (C.52) and (C.55) that

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{M}(L,\gamma,K)} \mathbb{P}\left\{\frac{1}{np}\|\widehat{M} - M\|_F^2 \geq A\left(\frac{n \vee p}{N}\right)^{\frac{2L}{2L+K}}\right\} \geq 1/2 \tag{C.56}$$

where $A = \frac{C_{2,L,K}\gamma^2}{16}(2c_0)^{-2L/K}$. With the selection of c_0 in (C.54), A depends on $L, K, \gamma, \kappa, \sigma^2$.

Thus, Theorem 4.10 is proved.

□

BIBLIOGRAPHY

[1] Aylin Alin. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):370–374, 2010.

[2] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Etats-Unis Mathématicien. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

[3] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[4] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[5] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[6] Francis R Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9(Jun):1019–1048, 2008.

[7] Rina Foygel Barber, Mladen Kolar, et al. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422–3450, 2018.

[8] Sascha O Becker, Andrea Ichino, et al. Estimation of average treatment effects based on propensity scores. *The stata journal*, 2(4):358–377, 2002.

[9] Tim Bedford and Roger M Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

[10] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[12] Seth I Berger and Ravi Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 25(19):2466–2472, 2009.

[13] Jinbo Bi, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3(Mar):1229–1243, 2003.

[14] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[15] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *cell*, 122(6):947–956, 2005.

[16] Florentina Bunea, Yiyuan She, Hernando Ombao, Assawin Gongvatana, Kate Devlin, and Ronald Cohen. Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55(4):1519–1527, 2011.

[17] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[18] T Tony Cai, Wen-Xin Zhou, et al. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

[19] Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[20] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.

[21] E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010.

[22] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[23] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[24] Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[25] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

[26] Alexander L. Chistov and Dima Grigoriev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *MFCS*, 1984.

[27] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[28] Jen-hwa Chu, Scott T Weiss, Vincent J Carey, and Benjamin A Raby. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC systems biology*, 3(1):55, 2009.

[29] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[30] Péter Csermely, Vilmos Agoston, and Sandor Pongor. The efficiency of multi-target drugs: the network approach might help drug design. *Trends in pharmacological sciences*, 26(4):178–182, 2005.

[31] Cun-Hui, Zhang, Stephanie, S., and Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models.

[32] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[33] Stephen G Donald and Whitney K Newey. Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50(1):30–40, 1994.

[34] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[35] Jeff Douglas, Hae Rim Kim, Brian Habing, and Furong Gao. Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, 23(2):129–151, 1998.

[36] E. Elhamifar. High-rank matrix completion and clustering under self-expressive models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 73–81, USA, 2016. Curran Associates Inc.

[37] Jicong Fan and Jieyu Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34 – 41, 2018.

[38] Jicong Fan and Tommy W.S. Chow. Non-linear matrix completion. *Pattern Recognition*, 77:378–394, 5 2018.

[39] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

[40] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[41] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

[42] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[43] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[44] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

[45] Sara Van De Geer. *Estimation and testing under sparsity*. 2016.

[46] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[47] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

[48] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, and Richard A Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Biocomputing 2001*, pages 422–433. World Scientific, 2000.

[49] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[50] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[51] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[52] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[53] Sui Huang and Donald E Ingber. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Experimental cell research*, 261(1):91–103, 2000.

[54] Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.

[55] William G Jacoby. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613, 2000.

[56] Martin Jaggi and Marek Sulovskỳ. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.

[57] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational

deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[58] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer, 2016.

[59] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

[60] Olga Klopp et al. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[61] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 10 2011.

[62] Mario Köppen. The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, volume 1, pages 4–8, 2000.

[63] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[64] Arun K Kuchibhotla, Lawrence D Brown, and Andreas Buja. Model-free study of ordinary least squares linear regression. *arXiv preprint arXiv:1809.10538*, 2018.

[65] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[66] M. Laurent. *Matrix completion problems*, pages 221–229. Kluwer Academic Publishers, Netherlands, 2001. Pagination: 9.

[67] Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, 2019.

[68] Hua Liang and Runze Li. Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248, 2009.

[69] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[70] Visit Limsombunchai. House price prediction: hedonic price model vs. artificial neural network. In *New Zealand agricultural and resource economics society conference*, pages 25–26, 2004.

[71] Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. In *Advances in neural information processing systems*, pages 800–808, 2012.

[72] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2010.

[73] Shisong Ma, Qingqiu Gong, and Hans J Bohnert. An arabidopsis gene network based on the graphical gaussian model. *Genome research*, 17(11):000–000, 2007.

[74] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England journal of medicine*, 363(2):166–176, 2010.

[75] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[76] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[77] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[78] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[79] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

[80] Jerzy Neyman and Egon S Pearson. *On the use and interpretation of certain test criteria for purposes of statistical inference. Part I.* University of California Press, 2020.

[81] Carl M. O'Brien. Statistical learning with sparsity: The lasso and generalizations. *International Statistical Review*, 84(1):156–157, 2016.

[82] Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

[83] Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert systems with applications*, 42(6):2928–2934, 2015.

[84] Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, December 2011.

[85] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, August 2010.

[86] James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability*

and statistics: essays in honor of David A. Freedman, pages 335–421. Institute of Mathematical Statistics, 2008.

[87] Natali Ruchansky, Mark Crovella, and Evimaria Terzi. Targeted matrix completion. In *SDM*, 2017.

[88] Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *arXiv preprint arXiv:1702.05037*, 2017.

[89] Mari Dominique Drouet Kotz Samuel, Dominique Drouet Mari, and Samuel Kotz. *Correlation and dependence*. World Scientific, 2001.

[90] Van De Geer Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

[91] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[92] Elizabeth D Schifano, Lin Li, David C Christiani, and Xihong Lin. Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*, 92(5):744–759, 2013.

[93] Jacob Schreiber, Timothy J Durham, Jeffrey Bilmes, and William Stafford Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv*, 2018.

[94] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[95] Burkhardt Seifert and Theo Gasser. Local polynomial smoothing. *Encyclopedia of statistical sciences*, 7, 2004.

[96] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in neural information processing systems*, pages 2951–2961, 2017.

[97] Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

[98] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13:290–312, 1982.

[99] Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 1329–1336, Cambridge, MA, USA, 2004. MIT Press.

[100] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[101] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.

[102] Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning*, pages 855–862. ACM, 2007.

[103] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[104] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[105] Hiroyuki Toh and Katsuhisa Horimoto. Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics*, 18(2):287–297, 2002.

[106] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[107] Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats. 2009.

[108] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[109] Caroline Uhler. Gaussian graphical models: An algebraic and geometric perspective. *arXiv preprint arXiv:1707.04345*, 2017.

[110] Sara Van de Geer. Estimation and testing under sparsity. *Lecture notes in mathematics*, 2159, 2016.

[111] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[112] Mark J Van der Laan. *Multiple Testing Procedures with Applications to Genomics. Springer Series in Statistics*. Springer, 2008.

[113] Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[114] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[115] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[116] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[117] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

[118] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274. ACM, 2015.

[119] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

[120] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[121] Nanny Wermuth and Steffen Lilholt Lauritzen. On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 21–50, 1990.

[122] Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463. Association for Computational Linguistics, 2017.

[123] Brian D Williamson, Peter B Gilbert, Noah Simon, and Marco Carone. Nonparametric variable importance assessment using machine learning techniques. 2017.

[124] G. Xia, H. Sun, B. Chen, Q. Liu, L. Feng, G. Zhang, and R. Hang. Nonlinear low-rank matrix completion for human motion recovery. *IEEE Transactions on Image Processing*, 27(6):3011–3024, June 2018.

[125] Yunhua Xiang and Noah Simon. A flexible framework for nonparametric graphical modeling that accommodates machine learning. In *International Conference on Machine Learning*, pages 10442–10451. PMLR, 2020.

[126] Huiliang Xie, Jian Huang, et al. Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696, 2009.

[127] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[128] Wenchao Yu, Guangxiang Zeng, Ping Luo, Fuzhen Zhuang, Qing He, and Zhongzhi Shi. Embedding with autoencoder regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 208–223. Springer, 2013.

[129] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[130] Laura A Zager and George C Verghese. Graph similarity scoring and matching. *Applied mathematics letters*, 21(1):86–94, 2008.

[131] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[132] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[133] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[134] Ying Zhu, Zhuqing Yu, and Guang Cheng. High dimensional inference in partially linear models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2760–2769. PMLR, 2019.

[135] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.