Investigations in Ultra High-Dimensional Models

by

Michael Law

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Ya'acov Ritov, Chair
Professor Moulinath Banerjee
Professor Fred Feinberg
Professor Ambuj Tewari

Michael Law

mmylaw@umich.edu

ORCID iD: 0000-0003-2013-4818

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my PhD advisor Ya'acov Ritov for all of his wisdom, patience, and support during these past several years. Ya'acov is an exceptional scholar, who not only taught me statistics, but also topics as diverse as engineering and etymology among others. In addition, I would also like to thank Professors Mouli Banerjee, Liza Levina, Ambuj Tewari, Stilian Stoev, Ji Zhu, and Ziwei Zhu for all their helpful advice and insightful discussions.

Finally, I want to thank my family for all their love and support over the years.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . ii

LIST OF FIGURES . . . vi

LIST OF TABLES . . . vii

LIST OF APPENDICES . . . viii

ABSTRACT . . . ix

CHAPTER

1 Introduction . . . 1

2 Inference Without Compatibility: Using Exponential Weighting for Inference on a Parameter of a Linear Model . . . 4

- 2.1 Introduction . . . 4
  - 2.1.1 Organization of the Chapter . . . 6
  - 2.1.2 General Notation and Definitions . . . 6
- 2.2 Inference for $\boldsymbol{\beta}^*$ . . . 8
  - 2.2.1 The Special Case: $q = 1$ . . . 8
  - 2.2.2 Correlated Gaussian Errors . . . 13
  - 2.2.3 The General Case: $q > 1$ . . . 14
  - 2.2.4 Necessity of Sparsity Assumption . . . 15
- 2.3 Inference for σ_μ^2 and σ_ε^2 . . . 16
  - 2.3.1 Inference for σ_μ^2 . . . 17
  - 2.3.2 Inference for σ_ε^2 . . . 19
- 2.4 Implementation . . . 20
- 2.5 Simulations . . . 23
  - 2.5.1 Simulations for β^* . . . 23
  - 2.5.2 Simulations for σ_μ^2 and σ_ε^2 . . . 25
- 2.6 Proofs . . . 28
  - 2.6.1 Proofs for Section 2.2.1 . . . 28
  - 2.6.2 Proofs for Section 2.2.4 . . . 37

3 Inference and Estimation for Random Effects in High-Dimensional Linear Mixed Models . . . 39

3.1 Introduction . . . 39
3.1.1 Organization of the Chapter . . . 40
3.1.2 Notation . . . 41
3.2 Hypotheses Testing for Random Effects . . . 41
3.2.1 Model and Motivation . . . 42
3.2.2 Estimator . . . 43
3.2.3 Assumptions . . . 45
3.2.4 Main Results . . . 48
3.3 Confidence Intervals for a Single Random Effect . . . 50
3.3.1 Model and Motivation . . . 50
3.3.2 Estimator . . . 51
3.3.3 Assumptions . . . 52
3.3.4 Main Results . . . 53
3.4 Empirical Bayes in ANOVA Type Models . . . 54
3.4.1 Model and Motivation . . . 55
3.4.2 Estimator . . . 55
3.4.3 Assumptions . . . 56
3.4.4 Main Results . . . 56
3.5 Simulations . . . 57
3.5.1 Methods and Models . . . 57
3.5.2 Results . . . 59
3.6 Real Data Application . . . 60

4 High-Dimensional Varying Coefficient Models with Functional Random Effects . . . 61

4.1 Introduction . . . 61
4.1.1 Organization of the Chapter . . . 63
4.1.2 Notation and Definitions . . . 63
4.2 Estimation with No Time Invariant Covariates . . . 65
4.2.1 Sample Size . . . 67
4.2.2 Assumptions . . . 68
4.2.3 Main Results . . . 69
4.3 Estimation with No Time Varying Covariates . . . 71
4.3.1 Assumptions . . . 72
4.3.2 Main Results: Independent Sampling Times . . . 74
4.3.3 Main Results: Common Sampling Times . . . 76
4.4 Two-Stage Estimation . . . 78
4.5 Confidence Bands . . . 79
4.5.1 Assumptions . . . 81
4.5.2 Main Results . . . 81
4.6 Simulations . . . 83
4.7 Human Height Data . . . 85

5 Rank-Constrained Least-Squares: Prediction and Inference . . . 88

5.1 Introduction . . . 88
5.1.1 Organization of the Chapter . . . 91

5.2 In-Sample Prediction Risk of the Rank-Constrained Estimator 91
5.3 Testing in Signal Plus Noise Models 96
5.3.1 A General Power Analysis of a Permutation Test 96
5.3.2 Sparse High-Dimensional Linear Model 100
5.3.3 Low-Rank Trace Regression 102
5.3.4 Robustness of the Rank-Constrained Test 104
5.4 Simulations 105
5.4.1 Models and Methods 105
5.4.2 In-Sample Prediction 106
5.4.3 Inference 106

APPENDICES 113

BIBLIOGRAPHY 184

LIST OF FIGURES

FIGURE

4.1 95% Confidence bands for $\boldsymbol{\beta}_1^*(t)$ 85
4.2 Estimated coefficient for the intercept, with North Africa as the reference region . . 87

5.1 Illustration of the construction of $\mathbf{X}_{\mathbf{V}^*}$ with oracle $\mathbf{V}^*$ when $r^* = 2$. 92
5.2 Plots of in-sample prediction error for Gaussian design 107
5.3 Plots of in-sample prediction error for matrix completion 107
5.4 Plots of power for Gaussian design 110
5.5 Plots of power for matrix completion 110

A.4.1 Marginal confidence bands for ABF1 and MAC1 169

LIST OF TABLES

TABLE

2.1 Simulations for β^* with Gaussian design and errors when $q = 1$ and $\beta^* = 0$ 25
2.2 Simulations for σ^2_μ with $s_\gamma = 3$ 27
2.3 Simulations for σ^2_ε with $s_\gamma = 3$ 28

5.1 Simulations for In-Sample Prediction Risk for Gaussian Design 108
5.2 Simulations for In-Sample Prediction Risk for Matrix Completion 109
5.3 Simulations for Inference for Gaussian Design 111
5.4 Simulations for Inference for Matrix Completion 112

A.2.1 Simulations for β^* with Gaussian design and errors when $q = 3$ and $\beta^* = 0$ 113
A.2.2 Simulations for β^* with Gaussian design and errors when $q = 1$ and $\beta^* = 1$ 114
A.2.3 Simulations for β^* with Gaussian design and errors when $q = 3$ and $\beta^* = 1$ 114
A.2.4 Simulations for β^* with double exponential design and errors when q=1 and $\beta^* = 0$. 115
A.2.5 Simulations for β^* with double exponential design and errors when $q = 3$ and $\beta^* = 0$ 115
A.2.6 Simulations for β^* with double exponential design and errors when $q = 1$ and $\beta^* = 1$ 116
A.2.7 Simulations for β^* with double exponential design and errors when $q = 3$ and $\beta^* = 1$ 116
A.2.8 Simulations for β^* with scaled t design and errors when q=1 and $\beta^* = 0$ 117
A.2.9 Simulations for β^* with scaled t design and errors when $q = 3$ and $\beta^* = 0$ 117
A.2.10 Simulations for β^* with scaled t design and errors when $q = 1$ and $\beta^* = 1$ 118
A.2.11 Simulations for β^* with scaled t design and errors when $q = 3$ and $\beta^* = 1$ 118
A.2.12 Simulations for σ^2_μ with $s_\gamma = 15$ 119
A.2.13 Simulations for σ^2_ε with $s_\gamma = 15$ 120

A.3.1 Simulations with $d = 0$ and $s = 3$ 142
A.3.2 Simulations with $d = 200$ and $s = 3$ 143
A.3.3 Simulations with Gaussian errors when $s = 15$ 144

A.4.1 Simulations for $\boldsymbol{\beta}(\cdot)$ with Trigonometric Basis 170
A.4.2 Simulations for $\boldsymbol{\beta}(\cdot)$ with B-Spline Basis 170
A.4.3 Simulations for $\gamma(\cdot)$ 171

LIST OF APPENDICES

1 Appendix of Chapter 2 . . . . **113**

2 Appendix of Chapter 3 . . . . **128**

3 Appendix of Chapter 4 . . . . **145**

4 Appendix of Chapter 5 . . . . **172**

ABSTRACT

This dissertation considers the problem of estimation and inference in four high-dimensional models: (i) high-dimensional linear models, (ii) high-dimensional linear mixed effects models, (iii) high-dimensional varying coefficient models with functional random effects, and (iv) low-rank trace regression models. In the context of linear models, we propose procedures to construct asymptotic confidence intervals for low-dimensional parameters in the presence of high-dimensional nuisance covariates without the compatibility condition. Then, for linear mixed effects models, we consider a high-dimensional analogue of the Wald test for random effects, establishing its asymptotic distribution and power. In addition, we show that empirical Bayes estimation performs as well as the oracle asymptotically in estimating a part of the mean vector. Next, we consider a high-dimensional varying coefficient model with functional random effects. Under sampling times that are either fixed and common or random and independent, we propose a projection procedure to estimate and construct confidence bands for the varying coefficients. Finally, in low-rank trace regression, we establish an in-sample prediction error bound for the rank-constrained least-squares estimator and consider a permutation test for the entire matrix of regression coefficients.

CHAPTER 1

Introduction

In the modern era, we increasingly encounter high-dimensional datasets, where the number of potential explanatory variables exceeds our number of subjects. For example, in the social sciences, we may be interested in the effect of many socio-economic factors. Similarly, in large genetic studies, we may be interested in how individual genomes affect the underlying disease status. While these modern datasets have the potential to allow scientists to simultaneously control for many sources of variability, it necessitates the development of new statistical techniques when the tools of classical statistics are no longer applicable.

In this dissertation, we contribute to the growing literature on high-dimensional statistics in four aspects:

1. **Linear models:** As the prototypical example of a high-dimensional model, consider a linear model given by

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_2 + \varepsilon_i, \tag{1.0.1}$$

   where $i = 1, \ldots, n$ and $\mathbf{x}_i \in \mathbb{R}^p$ is the vector of covariates. Here, y_i is the response and ε_i is independent and identically distributed noise.

   In classical statistics, when $p < n$, the performance of the least-squares estimator is well understood; least-squares is the best linear unbiased estimator and semi-parametrically efficient when the errors are Gaussian. However, when $p > n$, the coefficient vector $\boldsymbol{\beta}^*$ is no longer identified and we need additional assumptions in order to conduct estimation and inference for $\boldsymbol{\beta}^*$. The most common assumption is coordinatewise sparsity, where $s^* \triangleq \|\boldsymbol{\beta}^*\|_0 < n$. The seminal work of Tibshirani (1996) introduced the lasso estimator, a popular regularized estimator that induces a sparse estimate of $\boldsymbol{\beta}^*$. Ten years elapsed before Candes and Tao (2007) provided the first statistical guarantees for the Dantzig selector, another regularized estimator for $\boldsymbol{\beta}^*$, which was then extended to the lasso estimator by Bickel et al. (2009).

   Later, Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang

(2014) provided the first guarantees to obtain an asymptotic distribution on a modified version of the lasso estimator, enabling statistical inference for low-dimensional parameters in the high-dimensional context. Despite the theoretical breakthrough, they required a technical condition known as the compatibility condition to ensure the validity of their estimator. In Chapter 2, we consider the problem of inference for low-dimensional parameters in a high-dimensional linear model without the technical compatibility condition. This chapter is published as Law and Ritov (2021b).

2. **Linear mixed effects models:** In many practical applications, our observations naturally exhibit structured dependence. A canonical example arises in the design of experiments, where each subject is randomly assigned to one of many treatment groups; in such a situation, we expect individuals who received the same treatment to be dependent. Similarly, it is common practice in medical settings to perform longitudinal studies, where individuals are observed over a fixed time period. Then, it is not surprising that measurements from the same individual are highly correlated.

   The linear mixed effects model is a natural extension of the linear model to account for this type of dependence. Consider the linear mixed effects model given by

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_2 + \langle \mathbf{z}_i, \boldsymbol{\nu} \rangle_2 + \varepsilon_i \tag{1.0.2}$$

   where $i = 1, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$, and $\mathbf{z}_i \in \mathbb{R}^q$. Again, y_i is the response and ε_i is independent and identically distributed noise. Here, $\boldsymbol{\beta}^*$ is the vector of regression coefficients corresponding to the fixed effects with covariates $\mathbf{x}_i$ while $\boldsymbol{\nu}$ is the vector of regression coefficients corresponding to the random effects with covariates $\mathbf{z}_i$. In the context of a longitudinal design, $\mathbf{z}_i$ is a standard basis vector encoding the individual and $\boldsymbol{\nu}$ is a vector containing the individual heterogeneity. Compared to the linear model given in equation (1.0.1), there is an additional term $\langle \mathbf{z}_i, \boldsymbol{\nu} \rangle_2$ to characterizes the dependence structure in our observations.

   In Chapter 3, we consider the problem of inference and estimation of the vector $\boldsymbol{\nu}$ when $q < n < p$, where the fixed effects are high-dimensional but the random effects are low-dimensional. This chapter is published as Law and Ritov (2022).

3. **Varying coefficient models with functional random effects:** In the previous two models, we have implicitly assumed either (i) the data is collected within a short time horizon or (ii) the underlying data generating mechanism is static over time. Modeling average human height data across generations is a natural example where we do not expect this assumption to hold.

   The varying coefficient model is a further refinement of the linear model to account for temporal influence. Consider the varying coefficient model with functional random effects given by

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j})\rangle_2 + \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j})\rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}) \tag{1.0.3}$$

for $i = 1, \ldots, n$, $t_{i,j} \in (0,1)$, $\mathbf{x}_i \in \mathbb{R}^p$, and $\mathbf{z}_i : (0,1) \to \mathbb{R}^q$. Here, i corresponds to our experimental units and $t_{i,j}$ are the sampling times for individual i. Again, $y_i(\cdot)$ represents our response while $\varepsilon_i(\cdot)$ is independent and identically distributed noise. The quantities $\mathbf{x}_i$ and $\mathbf{z}_i(\cdot)$ represent our time invariant and time varying covariates respectively while $\xi_i(\cdot)$ denotes the functional random effect. Similar to the linear mixed effects model, by incorporating functional random effects, we can encapsulate the induced heterogenity arising from repeated measurements on the same experimental unit. Note that the covariate corresponding to the random effect is simply an indicator for the subject. Thus, the model in equation (1.0.3) is an extension of the model in equation (1.0.2) to allow for temporal dependence.

In Chapter 4, we consider the problem of estimation and inference for both $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$. This chapter is published as Law and Ritov (2021a).

4. **Low-rank trace regression models:** Finally, we consider the low-rank trace regression model, an extension of the linear model to accommodate matrix valued covariates, which is given by

$$y_i = \langle \mathbf{X}_i, \boldsymbol{\Theta}^*\rangle_{\text{HS}} + \varepsilon_i$$

for $i = 1, \ldots, n$, where $\langle \mathbf{X}_i, \boldsymbol{\Theta}^*\rangle_{\text{HS}} \triangleq \operatorname{tr}(\mathbf{X}_i^\top \boldsymbol{\Theta}^*)$. Here, y_i is the response, $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$, $\boldsymbol{\Theta}^*$ is a matrix of regression coefficients, and ε_i is independent and identically distributed noise. In the high-dimensional context, we have $n < d_1 d_2$. Compared to the linear model in equation (1.0.1), we do not assume that $\boldsymbol{\Theta}^*$ is coordinatewise sparse, but rather $\boldsymbol{\Theta}^*$ is low-rank, which preserves the matrix structure of the data.

The existing literature focuses on the nuclear norm regularized estimator, the analogue of the lasso estimator to the matrix setting. Like the lasso estimator, the nuclear norm regularized estimator requires a technical condition on the design matrix to ensure consistent prediction.

In Chapter 5, we consider the problem of prediction under no assumptions on the design matrix and inference for $\boldsymbol{\Theta}^*$ with independent and identically distributed observations. This chapter is published as Law et al. (2021).

CHAPTER 2

Inference Without Compatibility: Using Exponential Weighting for Inference on a Parameter of a Linear Model

2.1 Introduction

In the past decade, there has been much interest in high-dimensional linear models, particularly following the work of Tibshirani (1996). However, it was not until the past few years that there have been methods to construct confidence intervals and p-values for particular covariates in the model. Consider a high-dimensional partially linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

with $\mathbf{X} \in \mathbb{R}^{n\times q}$, and $\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$. In addition, we also observe covariates $\mathbf{Z} \in \mathbb{R}^{n\times p}$ such that $\boldsymbol{\mu} \approx \mathbf{Z}\boldsymbol{\gamma}^*$ for some sparse vector $\boldsymbol{\gamma}^* \in \mathbb{R}^p$ (see Section 2.1.2 for details). The vector $\boldsymbol{\mu}$ represents some underlying random nuisance parameters in the model that affect the response $\mathbf{y}$; the covariates $\mathbf{Z}$ allow us to control for these confounding factors. Regarding the size of each matrix, we assume that $q < n$ is fixed but $p > n$ is high-dimensional. Our goal is to construct a confidence region for the entire vector $\boldsymbol{\beta}^* \in \mathbb{R}^q$.

In recent years, there have been mainly two approaches to constructing confidence intervals in high-dimensional linear models. There have been approaches such as Lee et al. (2016), which construct conditional confidence intervals for $\boldsymbol{\beta}^*$ given that $\boldsymbol{\beta}^*$ was selected by a procedure, such as the lasso. Simultaneously, there has been work to construct unconditional confidence intervals for $\boldsymbol{\beta}^*$, where $\mathbf{X}$ is the a priori selected covariate of interest, such as Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014); the latter is also our focus. To avoid digressions, we do not elaborate on the former. A review of many of the current methods is available in Dezeure et al. (2015). Much of the existing literature relies on using a version of the de-sparsified lasso introduced simultaneously by Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014). The idea behind the existing approaches is to invert the KKT conditions of the lasso and perform nodewise lasso to approximate the inverse covariance matrix of the design, which attempts to correct the bias introduced by the lasso.

Since the lasso forms the basis for the procedure, certain assumptions must be made in order to ensure that the lasso enjoys the nice theoretical properties that have been developed over the past two decades. The paper by van de Geer and Bühlmann (2009) provides an overview of various assumptions that have been used to prove oracle inequalities for the lasso. These assumptions are a consequence of the fact the lasso is used rather than being needed for the statistical problem. In particular, for confidence intervals, van de Geer et al. (2014) assume that the compatibility condition holds for the Gram matrix, which is the weakest assumption from van de Geer and Bühlmann (2009), and is essentially a necessary assumption for the lasso to enjoy the fast rate (cf. Bellec (2018)). To quote the popular book by Bühlmann and van de Geer (2011), "In fact, a compatibility condition is nothing else than simply an assumption that makes our proof go through." However, this raises an important question on necessity: Is the compatibility condition necessary for constructing confidence intervals in high-dimensions?

The main contribution of this paper is proving that the compatibility condition or any of its variants is indeed not necessary for the statistical problem. To this end, we provide an estimator which does not require the compatibility condition but still attains the semi-parametric efficiency bound. Our assumption regarding sparsity is slightly stronger than the minimax rate required by Javanmard and Montanari (2018) since we allow a broader class of designs. In particular, we show that, in the absence of compatibility, the rate established by Javanmard and Montanari (2018) is not attainable and a stronger sparsity assumption is required.

To help clarify the connection between our notion of partially linear model and the high-dimensional linear models of the aforementioned works, we note that our model is many times written as a linear model $y = \langle \mathbf{x}, \boldsymbol{\beta}^* \rangle_2 + \langle \mathbf{z}, \boldsymbol{\gamma}^* \rangle_2 + \varepsilon$, reserving the notion of partially linear model to $y = \langle \mathbf{x}, \boldsymbol{\beta}^* \rangle_2 + \mu(t) + \varepsilon$ for some unknown smooth function $\mu(\cdot)$. We use the PLM terminology to emphasize that (i) $\langle \mathbf{z}, \boldsymbol{\gamma}^* \rangle_2$ is only an approximation, and (ii) $\mathbf{z}$ is a high-dimensional nuisance parameter, which plays the role of the nonparametric part of a semi-parametric model. For more details, see Remark 1.1 below.

There is also the recent work of Chernozhukov et al. (2018a), who consider the general problem of conducting inference on low-dimensional parameters with high-dimensional nuisance parameters. One application of their general theory is for high-dimensional partially linear models, which is also our problem of interest. A further discussion of their procedure is given in Remark 2.2.1 below.

As a consequence of our estimation procedure for $\boldsymbol{\beta}^*$, we are able to construct a $\sqrt{n}$-consistent estimator of the signal strength and the noise variance, which we denote by σ_μ^2 and σ_ε^2 respectively, also without the compatibility condition. The paper by Reid et al. (2016) provides a nice overview of different proposals for estimation of σ_ε^2 using the lasso. An early work in this direction is Fan et al. (2012), who construct asymptotic confidence intervals for σ_ε^2 under a sure screening property of the covariates; in the setting of the lasso, this requires a β-min condition. Dicker (2014) consider a similar problem of variance estimation using moment estimators that do not require sparsity of the underlying signal. However, they do not consider the ultra high-dimensional setting nor the problem of inference. Later, Janson et al. (2017) considered inference on the signal-to-noise ratio but the theory developed only applies to Gaussian designs. For the problem of inference for σ_μ^2, the work most similar with ours is Cai and Guo (2020), who consider a more general problem in the semi-supervised setting, but their results for the supervised framework require minimal non-zero eigenvalues on the covariance matrix. To this end, we construct estimators that attain asymptotic variances equal to that of the efficient estimator in low-dimensions.

For both problems, our approach involves using exponential weighting to aggregate over all models of a particular size. Prima facie, this is a computationally hard problem but can be well approximated in practice. To this end, we propose an algorithm inspired by Rigollet and Tsybakov (2011).

2.1.1 Organization of the Chapter

We end the current section with the notation that is used throughout the paper. In Section 2.2, we discuss the problem of conducting inference for low-dimensional $\boldsymbol{\beta}^*$ in the presence of a high-dimensional nuisance vector $\boldsymbol{\mu}$. The setting of univariate β^* is considered separately in Section 2.2.1 to motivate the general multivariate procedure of Section 2.2.3. We take a slight detour in Section 2.2.2 to consider inference when the errors are correlated. The section ends with a discussion on the necessity of the sparsity assumption in Section 2.2.4. Then, in Section 2.3.1 and Section 2.3.2, we consider the problems of inference for σ_μ^2 and σ_ε^2 respectively. In Section 2.4, we provide an overview of the computation of the estimators, which we apply in Section 2.5 for numerical simulations. The proofs for Sections 2.2.1 and 2.2.4 are provided in Section 2.6. Additional simulation tables and the proofs for the remaining results are available in Appendix 1.

2.1.2 General Notation and Definitions

Throughout, all of our variables (except $\boldsymbol{\beta}^*$) have a dependence on n, but when it should not cause confusion, this dependence is suppressed. For a general vector $\mathbf{a}$ and a matrix $\mathbf{A}$, $\mathbf{a}_j$ denotes the jth entry of $\mathbf{a}$, $\mathbf{A}_j$ the jth column of $\mathbf{A}$, and $\mathbf{A}^{(j)}$ the jth row of $\mathbf{A}$. Then, $\|\mathbf{a}\|_2$ denotes the standard Euclidean norm, with the dimension of the space being implicit from the vector, $\|\mathbf{a}\|_1$ the L_1-norm, and $\|\mathbf{a}\|_0$ the L_0-norm. Furthermore, $\|\mathbf{A}\|_2$ will denote the operator norm and $\|\mathbf{A}\|_{\mathrm{HS}}$ the Hilbert-Schmidt norm. If $\mathbf{A}$ is square, $\mathbf{A}^{-1}$ is to be interpreted in a generalized sense whenever the matrix $\mathbf{A}$ is rank deficient.

Before defining weak sparsity, we need to introduce some notation. For $u \in \mathbb{N}$, $\mathcal{M}_u$ will denote the collection of all models of $\mathbf{Z}$ of size u. That is,

$$\mathcal{M}_u \triangleq \{\mathbf{m} \subseteq \{1, \ldots, p\} : |\mathbf{m}| = u\}.$$

Then, for each $\mathbf{m} \in \mathcal{M}_u$, $\mathbf{Z_m}$ will denote the $n \times u$ sub-matrix of $\mathbf{Z}$ corresponding to the columns indexed by $\mathbf{m}$. Moreover, $\mathbf{P_m}$ will denote the projection onto the column space of $\mathbf{Z_m}$ and $\mathbf{P}_{\mathbf{m}}^{\perp}$ the projection onto the orthogonal complement. We can now state the definition of weak sparsity.

Definition 2.1.1. A sequence of vectors $\boldsymbol{\mu}$ is said to satisfy the *weak sparsity property relative to* $\mathbf{Z}$ with sparsity s at rate k if the set

$$\mathcal{S}_{\boldsymbol{\mu}} \triangleq \left\{\mathbf{m} \in \mathcal{M}_s : \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|^2 = o(k)\right\}$$

is non-empty. A set $\boldsymbol{S} \in \mathcal{S}_{\boldsymbol{\mu}}$ is said to be a *weakly sparse set* for the vector $\boldsymbol{\mu}$.

If the sequence of vectors $\boldsymbol{\mu}$ is random, then they satisfy the *weak sparsity property relative to* $\mathbf{Z}$ *in probability* with sparsity s at rate k if the set

$$\mathcal{S}_{\boldsymbol{\mu}} = \left\{\mathbf{m} \in \mathcal{M}_s : \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|^2 = o_{\mathbb{P}}(k)\right\}$$

is non-empty. A set $\boldsymbol{S} \in \mathcal{S}_{\boldsymbol{\mu}}$ is said to be a *weakly sparse set in probability* for the vector $\boldsymbol{\mu}$.

Remark. There are two distinctions to be made, between *strong* and *weak* sparsity on one hand, and between *weak sparsity* and *weak sparsity in probability*. The following examples may help to clarify these notions.

First, suppose that $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\gamma}^*$ for a sparse vector $\boldsymbol{\gamma}^* \in \mathbb{R}^p$ with support $\boldsymbol{S}$. We refer to this case as *strong sparsity* and is the commonly assumed setting in high-dimensional linear models (for example, van de Geer et al. (2014)). Since $\|\mathbf{P}_{\boldsymbol{S}}^{\perp}\boldsymbol{\mu}\|^2 = 0$, strong sparsity implies weak sparsity.

Second, consider a smooth function $\mu(\cdot) : \mathbb{R} \to \mathbb{R}$. This corresponds to a standard partially linear model, where $\mu(t)$ might denote a dependence of the mean on time. Let $\mathbf{Z}$ be a dictionary of basis functions, say, the harmonic or wavelet basis. Then, $\boldsymbol{\mu}$ may be well approximated by a linear combination of a few basis functions, with the remainder converging to zero, and weak sparsity holds.

Third, in random designs, it may be that with small probability $\boldsymbol{\mu}$ is not well approximated by any members of $\mathcal{M}_s$, but only holds with high probability. This case is referred to as weak sparsity in probability.

In general, if $\mathcal{S}_\mu$ is non-empty, then we may let $\boldsymbol{\gamma}^* = (\mathbf{Z}_{\boldsymbol{S}}^\mathsf{T}\mathbf{Z}_{\boldsymbol{S}})^{-1}\mathbf{Z}_{\boldsymbol{S}}\boldsymbol{\mu}$ for any $\boldsymbol{S} \in \mathcal{S}_\mu$. Depending on context, we either view $\boldsymbol{\gamma}^*$ as a vector in $\mathbb{R}^p$ or $\mathbb{R}^s$.

Finally, similar to other works on de-biased inference, we will consider sub-Gaussian errors, which is defined below.

Definition 2.1.2. A mean zero random vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is said to be *sub-Gaussian* with parameter K if

$$\mathbb{E}\exp\left(\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\xi}\right) \leq \exp\left(\frac{K^2\left\|\boldsymbol{\lambda}\right\|_2^2}{2}\right)$$

for all vectors $\boldsymbol{\lambda} \in \mathbb{R}^n$.

2.2 Inference for $\boldsymbol{\beta}^*$

In this section, we consider the main problem of constructing confidence regions for β^*. The model that we consider is given in equation (2.1.1), which we reproduce below for convenience,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \tag{2.2.1}$$

We write $\sigma_\varepsilon^2 \triangleq \mathrm{Var}(\varepsilon_1)$. For this section, we assume that $\boldsymbol{\mu}$ satisfies the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$, but the results still hold if we assume the weak sparsity property in probability.

2.2.1 The Special Case: $q = 1$

In this sub-section, we assume throughout that $q = 1$. In addition to the partially linear model given in equation (2.2.1), we also assume that $\mathbf{x}$satisfies a partially linear model, denoted by

$$\mathbf{x} = \boldsymbol{\nu} + \boldsymbol{\eta}, \tag{2.2.2}$$

where $\boldsymbol{\nu}$ satisfies the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$. We allow the weakly sparse set for $\boldsymbol{\nu}$ to be different from that of $\boldsymbol{\mu}$. We also assume that $\boldsymbol{\eta}$ is a sub-Gaussian vector with variance $\sigma_\eta^2 \triangleq \mathrm{Var}(\eta_1)$. The sub-Gaussianity assumption is needed to ensure that the empirical estimate of the norm squared residuals approximates the expectation well enough. By direct substitution, it follows that

$$\mathbf{y} = \boldsymbol{\nu}\beta^* + \boldsymbol{\mu} + \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}.$$

Since $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ both satisfy the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$, the vector $\boldsymbol{\nu}\beta^* + \boldsymbol{\mu}$ also satisfies the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$. To motivate our procedure, we assume temporarily that the models are in fact low-dimensional linear models. That is, suppose there are sets $\boldsymbol{S}_\delta$ and $\boldsymbol{S}_\gamma$ such that $\boldsymbol{\nu} = \mathbf{Z}_{\boldsymbol{S}_\delta}\boldsymbol{\delta}^*$ and $\boldsymbol{\mu} = \mathbf{Z}_{\boldsymbol{S}_\gamma}\boldsymbol{\gamma}^*$ for sparse vectors $\boldsymbol{\delta}^*$ and $\boldsymbol{\gamma}^*$. Moreover, assume that the set $\boldsymbol{S} \triangleq \boldsymbol{S}_\delta \cup \boldsymbol{S}_\gamma$ is known and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0_n, \sigma_\varepsilon^2 I_n\right)$. Thus, we are temporarily assuming the low-dimensional linear models

$$\begin{aligned}
\mathbf{y} &= \mathbf{x}\beta^* + \mathbf{Z}_{\boldsymbol{S}_\gamma}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon} = \mathbf{Z}_{\boldsymbol{S}\boldsymbol{\theta}+\boldsymbol{\eta}\beta^*+\boldsymbol{\varepsilon}}, \\
\mathbf{x} &= \mathbf{Z}_{\boldsymbol{S}_\delta}\boldsymbol{\delta}^* + \boldsymbol{\eta},
\end{aligned}$$

where $\boldsymbol{\theta}^* = \boldsymbol{\delta}^*\beta^* + \boldsymbol{\gamma}^*$. Then, by the Gauss-Markov Theorem, it is known that the efficient estimator in this low-dimensional problem is given by least-squares, which may be framed as the following three stage procedure:

1. Regress $\mathbf{y}$ on $\mathbf{Z}_{\boldsymbol{S}}$ using least-squares to obtain the fitted values $\hat{\mathbf{y}}$.
2. Regress $\mathbf{x}$ on $\mathbf{Z}_{\boldsymbol{S}}$ using least-squares to obtain the fitted values $\hat{\mathbf{x}}$.
3. Regress the residuals $\mathbf{y} - \hat{\mathbf{y}}$ on the the residuals $\mathbf{x} - \hat{\mathbf{x}}$ using least-squares to obtain the least-squares estimator $\hat{\beta}_{\text{LS}}$.

In the high-dimensional setting, the first two stages can no longer be achieved using the classical least-squares approach. However, since we are only interested in the fitted values $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$, this suggests using a high-dimensional prediction procedure to obtain the fitted values, and then applying low-dimensional least-squares on the residuals in the third stage. The high-dimensional procedure that we adopt is the exponential weights of Leung and Barron (2006), which has the salient feature of prediction consistency under very mild assumptions on the design.

Before defining our estimators, we state all of our assumptions.

(2.1) The means $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ have squared norms that are $\mathcal{O}_{\mathbb{P}}(n)$.

(2.2) The entries of $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are mutually independent and also independent of $\mathbf{Z}$. Moreover, the entries of $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are each identically distributed sub-Gaussians with parameters K_η and K_ε respectively.

(2.3) The means $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, and $\boldsymbol{\nu}\beta^* + \boldsymbol{\mu}$ are weakly sparse relative to $\mathbf{Z}$ with sparsities s_γ, s_δ, and s_θ respectively at rate $\sqrt{n}$. Furthermore, it is assumed that the statistician knows sequences u_γ, u_δ and u_θ with $u_\gamma \geq s_\gamma$, $u_\delta \geq s_\delta$, and $u_\theta \geq s_\theta$ for n sufficiently large and $\max\left(u_\gamma, u_\delta, u_\theta\right) = o(\sqrt{n}/\log(p))$.

Condition (2.1) ensures that the trivial situations in which the signal to noise ratios, $\|\boldsymbol{\mu}\|_2^2/n\sigma_\varepsilon^2$ and $\|\boldsymbol{\nu}\|_2^2/n\sigma_\eta^2$, respectively, are bounded away from zero and infinity asymptotically. Now, we may define two sets of exponential weights, $w_{\mathbf{m},y}$ and $w_{\mathbf{m},x}$, to estimate $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ respectively. Let

$$w_{\mathbf{m},y} \triangleq \frac{\exp\left(-\frac{1}{\alpha_y}\left\|\mathbf{P}_{\mathbf{m}}^{\perp}\mathbf{y}\right\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_{u_\theta}}\exp\left(-\frac{1}{\alpha_y}\left\|\mathbf{P}_{\mathbf{k}}^{\perp}\mathbf{y}\right\|_2^2\right)}$$

with $\alpha_y > 4K_\varepsilon^2$.

Remark. The exponential weights defined above do not subtract off the rank of the projection in the exponent as in Leung and Barron (2006) since we only consider models of size u_θ; the rank will cancel from the numerator and the denominator.

Now, let $\hat{\boldsymbol{\theta}}_{\mathbf{m}} \triangleq \left(\mathbf{Z}_{\mathbf{m}}^{\mathsf{T}}\mathbf{Z}_{\mathbf{m}}\right)^{-1}\mathbf{Z}_{\mathbf{m}}^{\mathsf{T}}\mathbf{y}$ be the least-squares estimator for $\boldsymbol{\theta}$ using the covariates $\mathbf{Z}_{\mathbf{m}}$. We identify $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ with a vector in $\mathbb{R}^p$, with the support of $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ being indexed by $\mathbf{m}$. Then, we may estimate $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}_{\text{EW}} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_{u_\theta}} w_{\mathbf{m},y}\hat{\boldsymbol{\theta}}_{\mathbf{m}},$$

with the prediction $\hat{\mathbf{y}}$ given by $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}$. Similarly, we define

$$w_{\mathbf{m},x} \triangleq \frac{\exp\left(-\frac{1}{\alpha_x}\left\|\mathbf{P}_{\mathbf{m}}^{\perp}\mathbf{x}\right\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_{u_\delta}}\exp\left(-\frac{1}{\alpha_x}\left\|\mathbf{P}_{\mathbf{k}}^{\perp}\mathbf{x}\right\|_2^2\right)},$$

with $\alpha_x > 4K_\eta^2$. Letting $\hat{\boldsymbol{\delta}}_{\mathbf{m}}$ denote the least-squares estimator of $\boldsymbol{\delta}^*$ using the covariates $\mathbf{Z}_{\mathbf{m}}$ and identifying it with a vector in $\mathbb{R}^p$, we may define

$$\hat{\boldsymbol{\delta}}_{\text{EW}} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_{u_\delta}} w_{\mathbf{m},\mathbf{x}}\hat{\boldsymbol{\delta}}_{\mathbf{m}}.$$

Then, the fitted values of $\mathbf{x}$ are $\hat{\mathbf{x}} = \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}$. Finally, for the last stage, the regression of $\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}$ on $\mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}$ is given by

$$\hat{\beta}_{\text{EW}} \triangleq \frac{\left(\mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right)}{\left\|\mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right\|_2^2}.$$

Before stating our main result, we state a proposition regarding exponential weighting with sub-Gaussian errors.

Proposition 2.1. *Consider a high-dimensional linear model given by*

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi},$$

for $\boldsymbol{\xi}$ sub-Gaussian with parameter K_ξ. Assume that $\boldsymbol{\mu}$ is weakly sparse relative to $\mathbf{Z}$ with sparsity s and that $\limsup_{n\to\infty} \|\boldsymbol{\mu}\|_2^2 = \mathcal{O}(n)$. Assume further that the chosen sequence of sparsities $u \geq s$ satisfy $u = o(n^\tau / \log(p))$ for $\tau > 0$ fixed. Letting $\hat{\boldsymbol{\gamma}}_{\mathbf{m}}$ denote the least-squares estimator for $\boldsymbol{\gamma}^*$ using the covariates $\mathbf{Z}_{\mathbf{m}}$, define the exponential weights as

$$w_{\mathbf{m}} \triangleq \frac{\exp\left(-\frac{1}{\alpha}\left\|\mathbf{P}_{\mathbf{m}}^{\perp}\mathbf{y}\right\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_u}\exp\left(-\frac{1}{\alpha}\left\|\mathbf{P}_{\mathbf{k}}^{\perp}\mathbf{y}\right\|_2^2\right)},$$

with $\alpha > 4K_\xi^2$. Then,

$$\mathbb{E}\left\|\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\mathbf{Z}\hat{\gamma}_{\mathbf{m}} - \boldsymbol{\mu}\right\|_2^2 = o(n^\tau).$$

Remark. We would like to remark that the choice of α is consistent with Leung and Barron (2006). In particular, when $\boldsymbol{\xi} \sim \mathcal{N}_n\left(0_n, \sigma_\xi^2 I_n\right)$, the sub-Gaussian parameter is $K^2 = \sigma_\xi^2$, which gives the requirement that $\alpha > 4\sigma_\xi^2$. In this setting, we emphasize that the required value of α is not consistent with a simple Bayesian interpretation since the Bayes procedure requires a leading constant of 2, as shown by Leung and Barron (2006). However, one of the referees pointed out that Grünwald and van Ommen (2017) show a way of explaining this in a Bayesian way in some extended models.

Remark. The assumption that $\limsup_{n\to\infty}\|\boldsymbol{\mu}\|_2^2 = \mathcal{O}(n)$ can be relaxed to hold in probability by weakening the conclusion to hold in probability rather than expectation (cf. Corollary 2.6.1).

For the remainder of the paper, we only consider the setting where $\tau = 1/2$. As an immediate corollary, we have the following.

Corollary 2.1.1. *Consider the models given in equations (2.2.1) and (2.2.2) with $q = 1$. Under assumptions (2.1) – (2.3),*

$$\left\|\boldsymbol{\nu}\beta^* + \boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{EW}\right\|_2^2 = o_{\mathbb{P}}(\sqrt{n}),$$
$$\left\|\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{EW}\right\|_2^2 = o_{\mathbb{P}}(\sqrt{n}).$$

Finally, we can state the main result for $\hat{\beta}_{\text{EW}}$.

Theorem 2.2. *Consider the models given in equations* (2.2.1) *and* (2.2.2) *with* $q = 1$. *Under assumptions (2.1) – (2.3),*

$$\sqrt{n}\left(\hat{\beta}_{EW} - \beta^*\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right).$$

We would like to note that $\hat{\beta}_{\text{EW}}$ attains the information bound for estimating β^* (cf. Example 2.4.5 of Bickel et al. (1993) and Section 2.3.3 of van de Geer et al. (2014)). Moreover, the convergence of $\hat{\beta}_{\text{EW}}$ is actually uniform. Consider the following parameter space

$$\mathcal{B} \triangleq \left\{(\beta^*, \sigma_\eta^2, \sigma_\varepsilon^2, K_\eta, K_\varepsilon) : \beta^* \in \mathbb{R}, \sigma_\eta^2 > 0, \sigma_\varepsilon^2 > 0, K_\eta > 0, K_\varepsilon > 0\right\}.$$

This induces a set of probability measures $(\mathscr{P}_\vartheta)_{\vartheta \in \mathcal{B}}$. Then, we have the following corollary.

Corollary 2.2.1. *Let* $\mathscr{K}$ *be a compact set of* $(\mathscr{P}_\vartheta)_{\vartheta \in \mathcal{B}}$ *with respect to variational distance. Under the setup of Theorem 2.2,*

$$\sqrt{n}\left(\hat{\beta}_{EW} - \beta^*\right) = a + b,$$

where

$$a \sim \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right),$$
$$|b| = o_{\mathbb{P}}(1)$$

uniformly for $\vartheta \in \mathscr{K}$.

Corollary 2.2.1 asserts that $\hat{\beta}_{\text{EW}}$ is uniformly Gaussian regular. Like Theorem 2.3 of van de Geer et al. (2014), the estimator $\hat{\beta}_{\text{EW}}$ is regular on one-dimensional parametric sub-models of (14) of van de Geer et al. (2014) and attains asymptotic semi-parametric efficiency. The main difference is replacing the assumption of compatibility of the design with the sparsity assumption (2.3).

Remark. The estimator, $\hat{\beta}_{\text{EW}}$, at first glance seems similar to the double/de-biased estimator of Chernozhukov et al. (2018a) by considering exponential weighting as the estimation procedure for the propensity function. However, the primary difference is that we do not rely on cross fitting to estimate the conditional mean of $\mathbf{x}$ and $\mathbf{y}$ given the covariates $\mathbf{Z}$. Therefore, $\hat{\beta}_{\text{EW}}$ does not fall within the general framework of Chernozhukov et al. (2018a) since we are using exponential weighting to solve in the in-sample prediction problem.

To construct confidence intervals, we need to estimate both σ_ε^2 and σ_η^2. We defer explicitly defining estimators for the variance until Section 2.3.2 but let $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\eta^2$ be any of the three estimators proposed by Theorem 2.10 for estimating variance. Then, an asymptotic $(1-\alpha)$ confidence interval for β^* is given by

$$\left(\hat{\beta}_{\mathrm{EW}} - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\eta^2 n}}, \hat{\beta}_{\mathrm{EW}} + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\eta^2 n}}\right),$$

where $z_{\alpha/2}$ denotes the $\alpha/2$ upper quantile of the standard Gaussian distribution.

2.2.2 Correlated Gaussian Errors

In this sub-section, we take a slight detour away from classical high-dimensional partially linear models and consider the setting where the errors, ε, are Gaussian but not necessarily independent and identically distributed. The goal is to conduct inference on β^*, but, for simplicity, we only consider the setting where $q = 1$. This model arises naturally if the model was a linear mixed model given by

$$\mathbf{y} = \mathbf{x}\beta^* + \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\zeta} + \boldsymbol{\xi},$$

where $\boldsymbol{\zeta}$ are Gaussian random effects and $\boldsymbol{\xi}$ is independent Gaussian noise. Bradic et al. (2019) and Li et al. (2019) consider more general problems of testing fixed effects in high-dimensional linear mixed models, whereas we simply view the problem as a linear model with correlated noise. Even when the errors are correlated, $\hat{\beta}_{\mathrm{EW}}$ still has a Gaussian limit under proper rescaling. Before stating the theorem, we will slightly modify assumption (2.2) to the setting where $\boldsymbol{\varepsilon}$ is correlated:

(2.2*) The entries of $\boldsymbol{\eta} \sim \mathcal{N}_n\left(0, \sigma_\eta^2 I_n\right)$ are independent of $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$. The vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \boldsymbol{\Sigma}_\varepsilon\right)$ is independent of $\mathbf{Z}$ with $\|\boldsymbol{\Sigma}_\varepsilon\| = \mathcal{O}(1)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_\varepsilon)/n \to \bar{d} > 0$.

Now, we may state the main result for $\hat{\beta}_{\mathrm{EW}}$ under correlation.

Theorem 2.3. *Consider the models given in equations* (2.2.1) *and* (2.2.2) *with* $q = 1$. *Under Assumptions (2.1), (2.2*), and (2.3),*

$$\sqrt{n}\left(\hat{\beta}_{EW} - \beta^*\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\bar{d}}{\sigma_\eta^2}\right).$$

Again, we defer defining an estimator for $\bar{d}$ and σ_η^2 until Section 2.3.2, in particular Corollary 2.3.2. Similar to the previous section, we may now construct confidence intervals for β^* under this setting of correlation.

2.2.3 The General Case: $q > 1$

In the general setting where $q > 1$, we may still rely on the perspective of high-dimensional prediction. Analogous to Section 2.2.1, we assume that each column of $\mathbf{X}$ satisfies a partially linear model. That is, there exist matrices $\mathbf{N}, \mathbf{H} \in \mathbb{R}^{n\times q}$ (read, capital $\mathbf{N}$ and capital $\mathbf{H}$, respectively) such that each column of $\mathbf{X}$ satisfies $\mathbf{X}_j = \mathbf{N}_j + \mathbf{H}_j$, where $\mathbf{N}_j$ satisfies the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$ for each $1 \leq j \leq q$. The weakly sparse set for each $\mathbf{N}_j$ may be different but the sparsity rate is uniformly $\sqrt{n}$. In matrix form, we have that

$$\mathbf{X} = \mathbf{N} + \mathbf{H}. \tag{2.2.3}$$

Since q is fixed and $\boldsymbol{\mu}$ and each $\mathbf{N}_j$ satisfy the weak sparsity property relative to Z at rate $\sqrt{n}$, the vector $\mathbf{N}\boldsymbol{\beta}^* + \boldsymbol{\mu}$ also satisfies the weak sparsity property relative to $\mathbf{Z}$ at rate $\sqrt{n}$. Moreover, $\mathbf{H}$ is assumed to be sub-Gaussian with the covariance matrix of each row of $\mathbf{H}$ given by $\boldsymbol{\Sigma}_H \triangleq \mathrm{Var}(\mathbf{H}^{(1)})$.

Then, for $1 \leq j \leq q$, we may let $\hat{\boldsymbol{\delta}}_{\mathrm{EW},j}$ denote the analogue of $\hat{\boldsymbol{\delta}}_{\mathrm{EW}}$ for regressing $\mathbf{X}_j$ on $\mathbf{Z}$ and estimate $\mathbf{X}_j$ by $\mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW},j}$. Let $\hat{\boldsymbol{\Delta}}_{\mathrm{EW}} \in \mathbb{R}^{p\times q}$ denote the matrix with columns given by $\hat{\boldsymbol{\delta}}_{\mathrm{EW},j}$ for $1 \leq j \leq q$. Then, the multidimensional analogue of $\hat{\boldsymbol{\beta}}_{\mathrm{EW}}$ from Section 2.2.1 is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{EW}} \triangleq \left(\left(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Delta}}_{\mathrm{EW}}\right)^{\mathsf{T}}\left(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Delta}}_{\mathrm{EW}}\right)\right)^{-1}\left(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Delta}}_{\mathrm{EW}}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\mathrm{EW}}\right).$$

We would like to emphasize that the definition here is identical to that given in Section 2.2.1 when $q = 1$.

Then, we will make the following assumptions.

(2.4) The mean vectors $\boldsymbol{\mu}$ and $\mathbf{N}_j$ for $1 \leq j \leq q$ have squared norms that are uniformly $\mathcal{O}_{\mathbb{P}}(n)$.

(2.5) The rows of $\mathbf{H}$ and the entries of $\boldsymbol{\varepsilon}$ are independent and also independent of $\mathbf{Z}$. Moreover, the entries of the rows of $\mathbf{H}$ and the entries of $\boldsymbol{\varepsilon}$ are each identically distributed sub-Gaussian with parameters $K_{\eta,j}$ and K_ε respectively. Furthermore, $\boldsymbol{\Sigma}_H$ is an invertible matrix.

(2.6) All the mean vectors $\boldsymbol{\mu}$, $\mathbf{N}_j$ for $1 \leq j \leq q$, and $\mathbf{N}\boldsymbol{\beta}^* + \boldsymbol{\mu}$ are weakly sparse relative to $\mathbf{Z}$ with sparsities s_γ, $s_{\delta,j}$ for $1 \leq j \leq q$, and s_θ respectively at rate $\sqrt{n}$. Furthermore, it is assumed that the statistician knows sequences u_γ, $u_{\delta,j}$, and u_θ with $u_\gamma \geq s_\gamma$, $u_{\delta,j} \geq s_{\delta,j}$ for $1 \leq j \leq q$ and $u_\theta \geq s_\theta$ for n sufficiently large and $\max\left(u_\gamma, \max_{1\leq j\leq q}\left(u_{\delta,j}\right), u_\theta\right) = o(\sqrt{n}/\log(p))$.

We can now state the asymptotic distribution for $\hat{\boldsymbol{\beta}}_{\mathrm{EW}}$.

Theorem 2.4. *Consider the models given in equations* (2.2.1) *and* (2.2.3). *Under assumptions (2.4) – (2.6),*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{EW} - \boldsymbol{\beta}^*\right) \xrightarrow{\mathcal{L}} \mathcal{N}_q\left(\mathbf{0}_q, \sigma_\varepsilon^2 \boldsymbol{\Sigma}_H^{-1}\right).$$

Similar to before, to construct confidence regions, we need to estimate $\boldsymbol{\Sigma}_H$. Therefore, we consider

$$\hat{\boldsymbol{\Sigma}}_H \triangleq \frac{1}{n}\left(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right).$$

This leads to the following proposition.

Proposition 2.5. *Consider the models given in equations* (2.2.1) *and* (2.2.3). *Under assumptions (2.4), (2.5), and (2.6),*

$$\hat{\boldsymbol{\Sigma}}_H \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}_H.$$

Then, an asymptotic $(1-\alpha)$ confidence region for β^* is given by

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^q : \frac{n}{\hat{\sigma}_\varepsilon^2}\left(\hat{\boldsymbol{\beta}}_{\text{EW}} - \boldsymbol{\beta}\right)^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_H \left(\hat{\boldsymbol{\beta}}_{\text{EW}} - \boldsymbol{\beta}\right) \leq \chi^2_{q,\alpha}\right\},$$

where $\chi^2_{q,\alpha}$ denotes the α upper quantile of a χ^2_q random variable.

2.2.4 Necessity of Sparsity Assumption

In Section 2.2.1, it was assumed that both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are weakly sparse with sparsity s_γ and s_δ respectively at rate $\sqrt{n}$ in order for $\hat{\boldsymbol{\beta}}_{\text{EW}}$ to have an asymptotic Gaussian distribution. For simplicity, in the ensuing discussion, we will only consider the case where $q=1$, that there exists an $\boldsymbol{S} \in \mathcal{S}_\mu$ such that $\left\|\mathbf{P}_{\boldsymbol{S}}^{\perp}\boldsymbol{\mu}\right\|_2^2 = 0$, and the design $(\mathbf{X}, \mathbf{Z})$ is fully Gaussian with population covariance matrix $\boldsymbol{\Sigma}$. That is, $\boldsymbol{\Sigma} = \text{Var}(\mathbf{X}_1, \mathbf{Z}^{(1)})$. We write $\boldsymbol{\Sigma}_{Z,Z}$ to denote the $p \times p$ sub-block of $\boldsymbol{\Sigma}$ corresponding to $\mathbf{Z}$. Letting $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, it follows that

$$s_\delta = \left|\{1 \leq j \leq p : \boldsymbol{\Omega}_{1,j} \neq 0\}\right|,$$

which is equivalent to s_Ω from Javanmard and Montanari (2018). Compared to the de-biased lasso, Javanmard and Montanari (2018) showed that, if $s_\gamma = o(n/\log^2(p))$ and $\min(s_\gamma, s_\delta) = o(\sqrt{n}/\log(p))$, then the de-biased lasso has an asymptotic Gaussian distribution. However, $\hat{\boldsymbol{\beta}}_{\text{EW}}$ is a valid estimator on a larger class of designs, in particular incompatible designs, and Theorem 2.6

formalizes this trade-off between sparsity and compatibility. Before stating the theorem, we will need to introduce a bit of notation regarding our parameter space Θ, which is defined as

$$\begin{aligned}\boldsymbol{\Theta}(s_\gamma, s_\delta) \triangleq \{\boldsymbol{\vartheta} = \left(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\Sigma}_{Z,Z}, \sigma_\eta^2, \sigma_\varepsilon^2\right) : \|\boldsymbol{\gamma}^*\|_0 \le s_\gamma, \|\boldsymbol{\delta}^*\|_0 \le s_\delta,\\ \max\left((\boldsymbol{\gamma}^*)^\intercal \boldsymbol{\Sigma}_{Z,Z}\boldsymbol{\gamma}^*, (\boldsymbol{\delta}^*)^\intercal \boldsymbol{\Sigma}_{Z,Z}\boldsymbol{\delta}^*, \sigma_\eta^2, \sigma_\varepsilon^2\right) = \mathcal{O}(1)\}.\end{aligned}$$

Theorem 2.6. *For $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}(s_\gamma, s_\delta)$, consider the following model*

$$\begin{aligned}&\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)} \overset{i.i.d}{\sim} \mathcal{N}_p\left(\mathbf{0}_p, \boldsymbol{\Sigma}_{Z,Z}\right),\\ &\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_n\right),\\ &\boldsymbol{\eta} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma_\eta^2 \mathbf{I}_n\right),\\ &\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon},\\ &\mathbf{X} = \mathbf{Z}\boldsymbol{\delta}^* + \boldsymbol{\eta}.\end{aligned}$$

Assume that either $s_\gamma = o(\sqrt{n}/\log(p))$ or $s_\delta = o(\sqrt{n}/\log(p))$. If there exists a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ for all $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}(s_\gamma, s_\delta)$, then both $s_\gamma = \mathcal{O}(\sqrt{n}/\log(p))$ and $s_\delta = \mathcal{O}(\sqrt{n}/\log(p))$.

In light of the results of Javanmard and Montanari (2018), to construct a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$, it must be the case that either $s_\gamma = o(\sqrt{n}/\log(p))$ or $s_\delta = o(\sqrt{n}/\log(p))$. The previous theorem implies that the other sparsity must satisfy $\mathcal{O}(\sqrt{n}/\log(p))$. Assumption (2.3) is only mildly stronger, requiring $\max(s_\gamma, s_\delta) = o(\sqrt{n}/\log(p))$.

Corollary 2.6.1. *For $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}(s_\gamma, s_\delta)$, consider the model in Theorem 2.6. If there exists $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ for all $\boldsymbol{\vartheta} \in (s_\gamma, s_\delta)$, then $\max(s_\gamma, s_\delta) = \mathcal{O}(\sqrt{n}/\log(p))$.*

2.3 Inference for σ_μ^2 and σ_ε^2

In this section, we consider the problem of conducting inference for both σ_μ^2 and σ_ε^2. Dicker (2014), Janson et al. (2017), and Cai and Guo (2020) provide interesting applications of both estimation and inference to which we refer the interested reader. The main model that we consider is slightly different than that considered in the previous section. Since we are not interested in the contribution of any particular covariate, we do not need to distinguish $\mathbf{X}$ from $\mathbf{Z}$. Hence, we set $q = 0$ and consider the following model,

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \tag{2.3.1}$$

Unlike Section 2.2, we view $\boldsymbol{\mu}$ as a random quantity, with $\sigma_\mu^2 \triangleq \mathrm{Var}(\boldsymbol{\mu}_1)$. Thus, σ_μ^2 can be viewed as the explained variation in the data using the covariates $\mathbf{Z}$. Throughout this section, $\boldsymbol{S}_\gamma$ will denote the weakly sparse set for $\boldsymbol{\mu}$ with sparsity s_γ. When constructing a $\sqrt{n}$-consistent estimator for σ_μ^2, the asymptotic distribution depends on the variance of $\boldsymbol{\mu}_1^2$, which we denote by $\kappa_\mu \triangleq \text{Var}\left(\boldsymbol{\mu}_1^2\right)$. Similarly, we need to let $\kappa_\varepsilon \triangleq \text{Var}\left(\boldsymbol{\varepsilon}_1^2\right)$ when constructing confidence intervals for σ_ε^2.

2.3.1 Inference for σ_μ^2

To motivate our high-dimensional procedure, we start by considering the low-dimensional setting. Letting $\boldsymbol{S}_\gamma$ denote a weakly sparse set for $\boldsymbol{\mu}$ relative to $\mathbf{Z}$ and identifying $\boldsymbol{\gamma}^*$ with a vector in $\mathbb{R}^{s_\gamma}$, we temporarily consider the linear model

$$\mathbf{y} = \mathbf{Z}_{\boldsymbol{S}_\gamma}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}. \tag{2.3.2}$$

The natural estimator for σ_μ^2 is given by $n^{-1}\left\|\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}\right\|_2^2$. The following proposition shows that this natural estimator is in fact efficient for estimating σ_μ^2 with Gaussian errors.

Proposition 2.7. *Consider the model given in equation* (2.3.2). *Assume that the design $\mathbf{Z}_{\boldsymbol{S}_\gamma}$ has full column rank and $s_\gamma < n$ is fixed. Then, the estimator $n^{-1}\left\|\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}\right\|_2^2$ is efficient for estimating σ_μ^2.*

From the Central Limit Theorem, it is immediate that

$$\sqrt{n}\left(n^{-1}\left\|\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}\right\|_2^2 - \sigma_\mu^2\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \kappa_\mu + 4\sigma_\mu^2\sigma_\varepsilon^2\right).$$

In the high-dimensional setting, there are three natural extensions of this low-dimensional efficient estimator using exponential weighting. The first idea is to view $\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}$ as the predicted values of $\mathbf{y}$ and directly use take the squared norm of the predicted values given by exponential weighting. For $\mathbf{m} \in \mathcal{M}_{u_\gamma}$, let $\hat{\boldsymbol{\gamma}}_m$ denote the least-squares estimator for $\boldsymbol{\gamma}^*$ using the covariates $\mathbf{Z}_\mathbf{m}$ and set

$$\hat{\boldsymbol{\mu}} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{m},y}\hat{\boldsymbol{\gamma}}_m,$$

where $w_{\mathbf{m},y}$ is defined in Section 2.2.1. Then, we may consider the estimator

$$\hat{\sigma}_{\mu,I}^2 \triangleq \frac{1}{n}\left\|\hat{\boldsymbol{\mu}}\right\|_2^2.$$

Alternatively, we may take the perspective that exponential weights concentrate well around the models with high predictive capacity, which would suggest aggregating the squared norms,

$$\hat{\sigma}_{\mu,II}^2 \triangleq \frac{1}{n}\sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{m},y}\left\|\mathbf{P}_\mathbf{m}\mathbf{y}\right\|_2^2.$$

The last estimator that we consider is inspired by the low-dimensional maximum likelihood estimator for σ_ε^2 and the fact that $\mathrm{Var}(\mathbf{y}_1) = \sigma_\mu^2 + \sigma_\varepsilon^2$:

$$\hat{\sigma}^2_{\mu,III} \triangleq \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|_2^2\right).$$

Before stating the main results for these estimators, we first provide all of our assumptions.

(2.8) The mean vector $\boldsymbol{\mu}$ has independent and identically distributed entries with finite fourth moment.

(2.9) The entries of $\boldsymbol{\varepsilon}$ are independent of $\mathbf{Z}$. Moreover, the entries of $\boldsymbol{\varepsilon}$ are independent and identically distributed sub-Gaussians with parameter K_ε.

(2.10) The vector $\boldsymbol{\mu}$ is weakly sparse relative to $\mathbf{Z}$ with sparsity s_γ. Furthermore, the chosen sparsity u_γ satisfies $u_\gamma = o(\sqrt{n}/\log(p))$ and $u_\gamma \geq s_\gamma$ for n sufficiently large.

Assumption (2.8) implies that $\|\boldsymbol{\mu}\|_2^2 = \mathcal{O}_{\mathbb{P}}(n)$. By Jensen's inequality, it is immediate that $\hat{\sigma}^2_{\mu,I} \leq \hat{\sigma}^2_{\mu,II} \leq \hat{\sigma}^2_{\mu,III}$. However, it turns out that, under the above assumptions, these estimators are asymptotically equivalent at the $\sqrt{n}$-rate. Recall that $\kappa_\mu \triangleq \mathrm{Var}(\boldsymbol{\mu}_1^2)$. The following theorem provides the asymptotic distribution of the three estimators.

Theorem 2.8. *Consider the model given in equation* (2.3.1). *Suppose that* $\sigma_\mu^2 > 0$. *Under assumptions (2.8) – (2.10),*

$$\sqrt{n}\left(\hat{\sigma}_\mu^2 - \sigma_\mu^2\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \kappa_\mu + 4\sigma_\varepsilon^2\sigma_\mu^2\right).$$

where $\hat{\sigma}_\mu^2$ *is either* $\hat{\sigma}^2_{\mu,I}$, $\hat{\sigma}^2_{\mu,II}$, *or* $\hat{\sigma}^2_{\mu,III}$.

Since our interest is mainly asymptotic, we write $\hat{\sigma}_\mu^2$ to denote generically one of the estimators for σ_μ^2. To construct confidence intervals for σ_μ^2, we need to estimate κ_μ, which may be accomplished by considering

$$\hat{\kappa}_\mu \triangleq \frac{1}{n}\sum_{j=1}^{n}\left(\hat{\boldsymbol{\mu}}_j^2 - \hat{\sigma}_\mu^2\right)^2.$$

The following proposition shows that $\hat{\kappa}_\mu$ is a consistent estimator for κ_μ.

Proposition 2.9. *Consider the model given in equation* (2.3.1). *Under assumptions (2.8) – (2.10),* $\hat{\kappa}_\mu \xrightarrow{\mathbb{P}} \kappa_\mu$.

Therefore, an asymptotic $(1-\alpha)$ confidence interval for σ_μ^2 is given by

$$\left(\hat{\sigma}_\mu^2 - z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\mu + 4\hat{\sigma}_\varepsilon^2\hat{\sigma}_\mu^2}{n}}, \hat{\sigma}_\mu^2 + z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\mu + 4\hat{\sigma}_\varepsilon^2\hat{\sigma}_\mu^2}{n}}\right). \tag{2.3.3}$$

2.3.2 Inference for σ_ε^2

In this section, we are interested in constructing confidence intervals for σ_ε^2. In the low-dimensional setting with Gaussian errors, an estimator for σ_ε^2 is given by maximum likelihood, which may be written as

$$\hat{\sigma}_{\varepsilon,\mathrm{ML}}^2 = \frac{1}{n}\left\|\mathbf{y} - \mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}\right\|_2^2.$$

From classical parametric theory, $\hat{\sigma}_{\varepsilon,\mathrm{ML}}^2$ is an efficient estimator for σ_ε^2 that achieves the information bound. A natural extension in the high-dimensional setting is to view $\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}$ as the predicted value and consider the estimator

$$\hat{\sigma}_{\varepsilon,I}^2 \triangleq \frac{1}{n}\left\|\mathbf{y} - \hat{\boldsymbol{\mu}}\right\|^2,$$

where $\hat{\boldsymbol{\mu}}$ is defined in Section 2.3.1. Recalling that $\mathrm{Var}\left(\mathbf{y}_1\right) = \sigma_\mu^2 + \sigma_\varepsilon^2$, we may consider two more estimators of σ_ε^2 in light of the results of Section 2.3.1, which are

1. $$\hat{\sigma}_{\varepsilon,II}^2 \triangleq \frac{1}{n}\left\|\mathbf{y}\right\|_2^2 - \hat{\sigma}_{\mu,II}^2.$$

2. $$\hat{\sigma}_{\varepsilon,III}^2 \triangleq \frac{1}{n}\left\|\mathbf{y}\right\|_2^2 - \hat{\sigma}_{\mu,I}^2.$$

Again, by Jensen's inequality, it is immediate that $\hat{\sigma}_{\varepsilon,I}^2 \leq \hat{\sigma}_{\varepsilon,II}^2 \leq \hat{\sigma}_{\varepsilon,III}^2$. Similar to before, these three estimators are asymptotically equivalent at the $\sqrt{n}$-rate and the following theorem provides the asymptotic distribution for all three.

Theorem 2.10. *Consider the model given in* (2.3.1) *with* $\sigma_\mu^2 > 0$. *Under assumptions (2.8) – (2.10),* $\sqrt{n}\left(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \kappa_\varepsilon\right)$, *where* $\hat{\sigma}_\varepsilon^2$ *is one of* $\hat{\sigma}_{\varepsilon,I}^2$, $\hat{\sigma}_{\varepsilon,II}^2$, *or* $\hat{\sigma}_{\varepsilon,III}^2$.

This gives us an immediate corollary to estimating $\bar{d}$ from Section 2.2.2, which requires the following assumption:

(2.2*) The vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon\right)$ is independent of $\mathbf{Z}$ with $\left\|\boldsymbol{\Sigma}_\varepsilon\right\|_2 = \mathcal{O}(1)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_\varepsilon)/n \to \bar{d} > 0$.

Consider the model given in equation (2.3.1). Under assumptions (2.8), (2.2*), and (2.10), $\hat{\sigma}^2_{\varepsilon,I} \xrightarrow{\mathbb{P}} \bar{d}$.

Remark. Currently, in this section, we have assumed that $q = 0$ but the theory for all three estimators of σ^2_ε are still valid when $q > 0$. In this setting, $\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\mu}$ is weakly sparse relative to $(\mathbf{X}, \mathbf{Z})$ with sparsity s_γ at rate $\sqrt{n}$. Therefore, by using exponential weighting with the design $(\mathbf{X}, \mathbf{Z})$, the above theorem implies that all three estimators are consistent for σ^2_ε.

Remark. In practice, one may consider a version of the three estimators dividing by $n - u_\gamma$ instead of n, consistent with the low-dimensional unbiased mean squared error estimator. Asymptotically, since $u_\gamma = o(\sqrt{n})$, they have the same asymptotic distribution but seem to have better performance empirically in finite sample.

Again, since $\hat{\sigma}^2_{\varepsilon,I}$, $\hat{\sigma}^2_{\varepsilon,II}$, and $\hat{\sigma}^2_{\varepsilon,III}$ are asymptotically equivalent, we write $\hat{\sigma}^2_\varepsilon$ to denote a generically any of the three estimators. To construct confidence intervals for σ^2_ε, we will need to estimate κ_ε. The estimator that we propose is similar to $\hat{\kappa}_\mu$, namely we will defined $\hat{\kappa}_\varepsilon$ as

$$\hat{\kappa}_\varepsilon \triangleq \frac{1}{n} \sum_{j=1}^{n} \left(\left(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_j \right)^2 - \hat{\sigma}^2_\varepsilon \right)^2 .$$

Analogous to Proposition 2.9, the following provides the consistency of $\hat{\kappa}_\varepsilon$.

Proposition 2.11. *Consider the model given in equation* (2.3.1)*. Under assumptions (2.8) – (2.10),* $\hat{\kappa}_\varepsilon \xrightarrow{\mathbb{P}} \kappa_\varepsilon$.

Therefore, an asymptotic $(1 - \alpha)$ confidence interval for σ^2_ε is given by

$$\left(\hat{\sigma}^2_\varepsilon - z_{\alpha/2} \sqrt{\frac{\hat{\kappa}_\varepsilon}{n}}, \hat{\sigma}^2_\varepsilon + z_{\alpha/2} \sqrt{\frac{\hat{\kappa}_\varepsilon}{n}} \right) . \tag{2.3.4}$$

2.4 Implementation

In this section, we describe a method to approximate all of the proposed estimators. Since all of our estimators are based on exponential weighting, we will only detail the task of estimating $\hat{\boldsymbol{\theta}}_{\text{EW}}$, with the others being analogous. Then, the goal of approximating $\hat{\boldsymbol{\theta}}_{\text{EW}}$ can be split into the following two tasks:

1. Determining the values of the tuning parameters α_y and u_θ.

2. Aggregating over $\binom{p}{u_\theta}$ models.

We start with the second task. Suppose temporarily that values of α_y and u_θ have been selected. To aggregate the models, we follow the Metropolis Hastings scheme of Rigollet and Tsybakov (2011). Our approach slightly differs from theirs since we restrict our attention to u_θ-sparse models whereas they consider models of varying sizes.

Conditional on the data, the values of $\hat{\boldsymbol{\theta}}_{\text{EW}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ for each $\mathbf{m} \in \mathcal{M}_{u_\theta}$ are fixed. We may view $\mathcal{M}_{u_\theta}$ as the vertices of the Johnson graph $J(p, u_\theta, u_\theta - 1)$ (cf. Godsil and Royle (2013)). Then, for each $\mathbf{m} \in \mathcal{M}_{u_\theta}$, by assigning weight $w_{\mathbf{m},y}$ to vertex $\mathbf{m}$, the target $\hat{\boldsymbol{\theta}}_{\text{EW}}$ may be viewed as the expectation of the fixed estimators $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ over the graph $J(p, u_\theta, u_\theta - 1)$, conditional on the observed data. Hence, by taking a random walk over $J(p, u_\theta, u_\theta - 1)$, we may approximate $\hat{\boldsymbol{\theta}}_{\text{EW}}$.

Before describing the algorithm, we need to introduce a bit of notation. For any model $\mathbf{m} \in \mathcal{M}_{u_\theta}$, we let $\mathcal{K}_{\mathbf{m}}$ denote the neighbors of $\mathbf{m}$, which is given by

$$\mathcal{K}_{\mathbf{m}} \triangleq \{\mathbf{k} \in \mathcal{M}_{u_\theta} : |\mathbf{k} \cap \mathbf{m}| = u_\theta - 1\}.$$

Moreover, we write $RSS_{\mathbf{m}} \triangleq \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\mathbf{y}\right\|_2^2$, the residual sum of squares. Furthermore, recall that if $\mathbf{Z}_{\mathbf{m}}^{\mathsf{T}}\mathbf{Z}_{\mathbf{m}}$ is rank deficient, then $(\mathbf{Z}_{\mathbf{m}}^{\mathsf{T}}\mathbf{Z}_{\mathbf{m}})^{-1}$ denotes any generalized inverse. Finally, let T_0 denote some burn-in time for the Markov chain and T denote the number of samples from the Markov chain. This yields the following algorithm, which closely parallels Rigollet and Tsybakov (2011).

Algorithm 1: Exponential weighting algorithm

Result: Approximates $\hat{\boldsymbol{\theta}}_{\text{EW}}$

Initialize a random point $\mathbf{m}_0 \in \mathcal{M}_{u_\theta}$ and compute $RSS_{\mathbf{m}_0}$;

for $t = 0, \dots, T$ **do**

Uniformly select $\mathbf{k} \in \mathcal{K}_{\mathbf{m}_t}$ and compute $RSS_{\mathbf{k}}$;

Generate a random variable $\mathbf{m}_{t+1}$ by

$$\mathbf{m}_{t+1} = \begin{cases} \mathbf{m}_t & \text{with probability } \exp\left(-\frac{1}{\alpha_y}(RSS_{\mathbf{k}} - RSS_{\mathbf{m}_t})\right); \\ \mathbf{k} & \text{with probability } 1 - \exp\left(-\frac{1}{\alpha_y}(RSS_{\mathbf{k}} - RSS_{\mathbf{m}_t})\right); \end{cases}$$

if $t > T_0$ **then**

Compute $\hat{\boldsymbol{\theta}}_{t+1} \leftarrow (\mathbf{Z}_{\mathbf{m}_{t+1}}^{\mathsf{T}} \mathbf{Z}_{\mathbf{m}_{t+1}})^{-1} \mathbf{Z}_{\mathbf{m}_{t+1}}^{\mathsf{T}} \mathbf{y}$, embedded as a vector in $\mathbb{R}^p$;

end

end

return

$$\frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \hat{\boldsymbol{\theta}}_{t+1};$$

Then, analogous to Theorem 7.1 of Rigollet and Tsybakov (2011), it will follow that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_{\text{EW}} \qquad \mathbb{P} \text{ almost surely.}$$

Now, for the first task, we may construct a grid of parameter points and use cross-validation to jointly tune the parameters using the above algorithm. Since both α_Y and u_θ do not need to be known exactly, but need to be tuned to be larger than a threshold, the grid can be quite coarse to ease the computational burden.

Computation in the ultrahigh-dimension is inherently difficult. In view of Zhang et al. (2014), there is no polynomial time algorithm that achieves the minimax rate for prediction without the restricted eigenvalue condition. However, we do not know any algorithm that verifies the restricted eigenvalue condition in polynomial time (cf Raskutti et al. (2010)). In this paper, we completely avoid assuming a condition like the restricted eigenvalue condition and therefore we cannot guarantee polynomial time convergence. Yet, the algorithm behaves well in practice, as can be seen from the simulations in the following section.

2.5 Simulations

We divide this section into two parts, corresponding to simulations for β^* and simulations for variance components σ_μ^2 and σ_ε^2. Additional simulation tables are included in the Supplement.

2.5.1 Simulations for β^*

For ease of comparison, our simulations will be similar to those given in van de Geer et al. (2014). For the linear models

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \\ \mathbf{X}_j &= \mathbf{N}_j + \mathbf{H}_j,\end{aligned}$$

we consider the setting where $n = 100$ and $p = 500$. There are a few parameters with which we experiment: q, $\boldsymbol{\beta}^*$, the distribution of the design and errors, the sparsities, and the signal to noise ratio. For each parameter pairing, we run 500 simulations. All confidence intervals are constructed at the nominal 95% level.

Since the number of parameters of interest is fixed and low-dimensional, we consider the settings where $q \in \{1, 3\}$. To assess both the coverage and the power, we let β^* be a vector in $\mathbb{R}^q$ with values in $\{0, 1\}$. To experiment with the robustness to the sub-Gaussianity assumption, we use Gaussian, double exponential, and $t(3)$ distributions for the errors, all scaled to have mean zero and unit variance. We denote these distributions by z, e, and t respectively. Therefore, $\sigma_\varepsilon^2 = 1$ throughout this section. The design have the same distribution as the error, but with an equi-correlation covariance matrix. That is, we consider the covariance matrix, $\boldsymbol{\Sigma}(Z)$ to be

$$\boldsymbol{\Sigma}(Z)_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \end{cases}$$

for $\rho \in \{0, 0.8\}$. When $q = 3$, the covariance matrix for $\mathbf{H}^{(1)}$, denoted by $\boldsymbol{\Sigma}(H)$, also is equi-correlation,

$$\boldsymbol{\Sigma}(H) = \begin{cases} \sigma_\eta^2 & \text{if } i = j. \\ 0.5\sigma_\eta^2 & \text{if } i \neq j, \end{cases}$$

where σ_η^2 is chosen so that $\mathrm{Var}(\mathbf{X}_1) = 1$.

Similar to van de Geer et al. (2014), we let the sparsity $s_\gamma \in \{3, 15\}$, and, for simplicity, set $s_\delta = s_\gamma$. We set the signal to noise ratio of $\boldsymbol{\mu}$ to $\boldsymbol{\varepsilon}$, which is given by $\sigma_\mu^2/\sigma_\varepsilon^2$, to be 2. Since large values of the signal to noise ratio (SNR) of $\mathbf{N}_j$ to $\mathbf{H}_j$ correspond to highly correlated designs, we also consider $SNR_X \triangleq \sigma_\nu^2/\sigma_\eta^2 \in \{2, 1000\}$.

For our simulations, we say $\boldsymbol{\mu}$ is weakly sparse relative to $\mathbf{Z}$ with sparsity s_γ at rate $\sqrt{n}$ if there exists an s_γ-sparse set S and vector γ_S such that $\text{Var}(\boldsymbol{\mu}_1 - (\mathbf{Z}_{\boldsymbol{S}}\boldsymbol{\gamma}_S^*)_1) \leq n^{-1/2}$. In particular, we consider vectors $\boldsymbol{\gamma}^*$ of the form

$$\boldsymbol{\gamma}_j^* \propto \pi(j)^{-\kappa} \qquad\qquad j = 1, \ldots, p$$

for some value $\kappa > 0$ and permutation $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$. A similar approach is applied for $\boldsymbol{\Delta}$.

We will compare our estimators with a few other procedures:

1. (LS) Oracle least-squares that knows the true weakly sparse set S_γ.

2. (DLA) De-biased lasso from Dezeure et al. (2015) as implemented in the R package `hdi`. We only apply this when $q = 1$.

3. (SILM) Simultaneous inference for high-dimensional linear models of Zhang and Cheng (2017) as implemented in the R package `SILM`.

4. (DML) Double/de-biased machine learning of Chernozhukov et al. (2018a) with 4 folds using the scaled lasso of Sun and Zhang (2012) as the estimation procedure as implemented in the R package `scalreg`. We only apply this when $q = 1$.

5. (EW$_I$), (EW$_{II}$), (EW$_{III}$) Exponential weights using $\hat{\sigma}^2_{\varepsilon,I}$, $\hat{\sigma}^2_{\varepsilon,II}$, and $\hat{\sigma}^2_{\varepsilon,III}$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

To evaluate the procedures, we use the following two measures

1. (AvgCov) Average coverage: The percentage of time the true value of β^* falls inside the confidence region.

2. (AvgLen) Average length: The average length of the confidence interval (only when $q = 1$).

The results are given in Table 1 and Tables A.2.1–A.2.11 from the Supplement. In the $q = 1$ setting with $SNR_X = 2$, the coverage is comparable amongst all of the estimators. However, the de-biased lasso and the SILM procedure are slightly preferable in this regime since the length of the intervals are slightly shorter. When $\beta^* = 0$, $SNR_X = 1000$, and $\rho = 0.8$, the coverage of the de-biased lasso is quite poor, with less than a 25% coverage against a nominal rate of 95%. The result should not be surprising since this corresponds to a setting of high correlation in the design, which weakens the compatibility condition. The double/de-biased machine learning approach has strong

Table 2.1: Simulations for β^* with Gaussian design and errors when $q = 1$ and $\beta^* = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.946	0.880	0.946	0.958	0.942	0.908	0.938	0.930
	DLA	0.958	0.884	0.976	0.978	0.954	0.870	0.218	0.170
	SILM	0.970	0.872	0.962	0.970	0.958	0.812	0.900	0.902
	DML	0.966	0.850	0.956	0.946	0.982	0.844	1.000	1.000
	EW_I	0.956	0.868	0.956	0.962	0.960	0.828	0.954	0.968
	EW_{II}	0.978	0.912	0.976	0.980	0.972	0.898	0.966	0.984
	EW_{III}	0.984	0.938	0.984	0.994	0.980	0.936	0.980	0.994
AvgLen	LS	0.427	0.462	0.589	0.684	0.430	0.467	0.919	1.440
	DLA	0.493	0.532	0.689	0.700	0.530	0.547	0.544	0.501
	SILM	0.529	0.559	0.670	0.697	0.623	0.609	0.666	0.646
	DML	0.650	0.634	0.694	0.692	1.510	0.881	10.600	11.100
	EW_I	0.623	0.636	0.700	0.716	1.060	0.774	1.910	1.830
	EW_{II}	0.690	0.710	0.768	0.797	1.170	0.868	2.100	2.040
	EW_{III}	0.749	0.776	0.830	0.871	1.280	0.951	2.270	2.240

nominal coverage in this regime (about 100%), but the length of the intervals are significantly longer than the other procedures (about four to five times longer than exponential weighting). When $\beta^* = 1$, $SNR_X = 1000$, and $\rho = 0.8$, we note that the SILM procedure no longer maintains nominal coverage. At first glance, it may seem odd that the oracle procedure based on least-squares does not always achieve the nominal coverage, but this is a consequence of weak sparsity. Since there is non-negligible bias in the model approximation in finite sample, this affects the empirical coverage of the oracle procedure. The results remain the same when we consider $q = 3$ and different distributions for the design and the errors. These results suggest that the compatibility assumption is crucial to the success of the lasso based procedures, and in the absence of such an assumption, the procedures based on exponential weighting maintain competitive coverage and length.

2.5.2 Simulations for σ_μ^2 and σ_ε^2

In this section, we set $q = 0$ and only consider the setting of strong sparsity (ie. $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\gamma}^*$ for some vector $\boldsymbol{\gamma}^* \in \mathbb{R}^p$ satisfying $\|\boldsymbol{\gamma}^*\|_0 = s_\gamma$). This reduces the linear model to $\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}$. We still consider the setting where $n = 100$ and $p = 500$. The value of $\sigma_\mu^2 = 2$ and $\sigma_\varepsilon^2 = 1$ throughout these simulations. The parameters with which we experiment are the distributions of the design and errors and the sparsity.

Again, we consider Gaussian, double exponential, and $t(3)$ distributions for the design and the errors. The design will have an equi-correlation structure with $\rho \in \{0, 0.8\}$ and the sparsity will satisfy $s_\gamma \in \{3, 15\}$.

The vector of coefficients, $\boldsymbol{\gamma}^*$, have s_γ components generated from uniform(-1,1) and $p - s_\gamma$ components that are zero. The values are then scaled such that $\sigma_\mu^2 = (\boldsymbol{\gamma}^*)^\top \boldsymbol{\Sigma}_Z \boldsymbol{\gamma}^* = 2$.

For estimation of σ_μ^2, we will compare our results with an oracular estimator based on low-dimensional least-squares and the recent proposal of CHIVE.

1. (LS) Oracle least-squares that knows the true strongly sparse set S_γ using equation (2.3.3).

2. (CHIVE) The calibrated inference for high-dimensional variance explained of Cai and Guo (2020). We follow Algorithm 1 of the paper with $\tau_0^2 \in \{0, 2, 4, 6\}$.

3. (EW$_I$), (EW$_{II}$), (EW$_{III}$) Exponential weighting using $\hat{\sigma}_{\mu,I}^2$, $\hat{\sigma}_{\mu,II}^2$, and $\hat{\sigma}_{\mu,III}^2$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

The results are presented in Table 2.2 and Table A.2.12 from the Supplement. We note that the coverage of the least-squares procedure is close to the nominal 95% rate when $s_\gamma = 3$ and the errors are either Gaussian or double exponential. The coverage is significantly worse for the $t(3)$ design, which should not be surprising since the fourth moment is not defined for this distribution. However, when $s_\gamma = 15$, the coverage of least-squares falls, which establishes a reference for the problem difficulty, since Proposition 2.7 establishes the efficiency of least-squares in this problem.

Amongst the exponential weighting estimators, when $s_\gamma = 3$ and the errors are Gaussian or double exponential, the procedure based on $\hat{\sigma}_{\mu,I}^2$ has the best performance and $\hat{\sigma}_{\mu,III}^2$ has the coverage when the errors are t distributed. For higher sparsity, no one estimators dominates the others; depending on our assumptions, any of the three estimators may be preferable. Compared with CHIVE, the best exponential weighting procedure seems to be able to achieve comparable coverage with significantly shorter intervals, which can be seen across all of our simulation settings.

Table 2.2: Simulations for σ_μ^2 with $s_\gamma = 3$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.922	0.948	0.914	0.934	0.808	0.802
	CHIVE$_0$	0.698	0.532	0.690	0.604	0.554	0.526
	CHIVE$_2$	0.818	0.668	0.792	0.702	0.712	0.634
	CHIVE$_4$	0.888	0.748	0.848	0.762	0.770	0.704
	CHIVE$_6$	0.890	0.772	0.898	0.790	0.860	0.746
	EW$_I$	0.852	0.850	0.854	0.862	0.780	0.778
	EW$_{II}$	0.804	0.772	0.820	0.838	0.812	0.828
	EW$_{III}$	0.708	0.644	0.744	0.762	0.820	0.866
AvgLen	LS	1.520	1.510	1.800	1.950	2.430	2.950
	CHIVE$_0$	0.998	0.937	1.160	1.190	1.670	2.130
	CHIVE$_2$	1.520	1.560	1.650	1.740	2.150	2.640
	CHIVE$_4$	1.890	1.970	2.010	2.120	2.500	2.980
	CHIVE$_6$	2.210	2.300	2.310	2.440	2.780	3.270
	EW$_I$	1.470	1.440	1.750	1.850	2.390	2.840
	EW$_{II}$	1.420	1.390	1.710	1.810	2.370	2.810
	EW$_{III}$	1.370	1.320	1.670	1.760	2.340	2.780

For the estimation of σ_ε^2, we will consider the oracular least-squares, the scaled lasso estimator, and the refitted cross-validation with Sure Independence Screening, along with our proposed procedures based on exponential weighting.

1. (LS) Oracle least-squares that knows the true strongly sparse set S_γ using equation (2.3.4).

2. (SL) Scaled lasso as implemented in the R package `scalreg` with a confidence interval constructed using Theorem 2 of Sun and Zhang (2012).

3. (RCV-SIS) Refitted cross-validation of Fan et al. (2012) using the Sure Independence Screening of Fan and Lv (2008) as implemented in the R package `SIS` in the first stage. The confidence interval is constructed using Theorem 2 of Fan et al. (2012), with $\mathbb{E}\varepsilon^4$ estimated by Proposition 2.11 of the present paper.

4. (EW$_I$), (EW$_{II}$), (EW$_{III}$) Exponential weighting using $\hat\sigma^2_{\varepsilon,I}$, $\hat\sigma^2_{\varepsilon,II}$, and $\hat\sigma^2_{\varepsilon,III}$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

The results are given in Table 2.3 and Table A.2.13 from the Supplement. When the signal is very sparse, $s_\gamma = 3$, and there is no correlation in the design, scaled lasso has better coverage than exponential weighting. However, as the correlation increases to $\rho = 0.8$, the confidence intervals constructed using $\hat{\sigma}^2_{\varepsilon,II}$ outperforms scaled lasso both in terms of coverage and average length. When the model is less sparse, $\hat{\sigma}^2_{\varepsilon,I}$ has comparable or better performance than scaled lasso. The poor performance of refitted cross-validation with Sure Independence Screening in the $s_\gamma = 15$ case should not come as a surprise since the signal to noise ratio is kept constant. The task of sure screening 15 active covariates out of 500 with low signal strength from 50 observations is very difficult.

Table 2.3: Simulations for σ^2_ε with $s_\gamma = 3$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.938	0.912	0.952	0.940	0.918	0.912
	SL	1.000	0.730	0.998	0.730	0.994	0.756
	RCV-SIS	0.684	0.646	0.688	0.644	0.638	0.606
	EW_I	0.616	0.608	0.678	0.674	0.650	0.690
	EW_{II}	0.862	0.828	0.872	0.846	0.852	0.814
	EW_{III}	0.672	0.458	0.660	0.488	0.636	0.430
AvgLen	LS	0.532	0.529	0.545	0.528	0.534	0.534
	SL	0.599	0.670	0.602	0.665	0.602	0.659
	RCV-SIS	0.485	0.509	0.508	0.514	0.554	0.539
	EW_I	0.430	0.427	0.442	0.438	0.435	0.447
	EW_{II}	0.441	0.444	0.453	0.453	0.446	0.463
	EW_{III}	0.462	0.475	0.473	0.480	0.466	0.492

2.6 Proofs

2.6.1 Proofs for Section 2.2.1

Before proving our main results, we state a simplified version of Theorem 2.1 of Hsu et al. (2012) will be useful in the subsequent proofs.

Lemma 2.12 (Theorem 2.1 of Hsu et al. (2012))**.** *Let* $\mathbf{P} \in \mathbb{R}^{n\times n}$ *be a rank* u *projection matrix. Let*

$\boldsymbol{\xi} \in \mathbb{R}^n$ *be a mean zero sub-Gaussian vector with parameter* K_ξ*. Then, for all* $t > 0$*,*

$$\mathbb{P}\left(\|\mathbf{P}\boldsymbol{\xi}\|_2^2 > K_\xi^2\left(u + 2\sqrt{ut} + 2t\right)\right) \leq \exp(-t).$$

For ease of reference in later proofs, we prove Proposition 2.1 as two lemmata.

Lemma 2.13. *Let* $\{w_{\mathbf{m}} : w_{\mathbf{m}} \geq 0, \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} = 1, \mathbf{m} \in \mathcal{M}_u\}$ *be weights, possibly random, and* ξ *be a sub-Gaussian vector with parameter* K_ξ*, independent of* $\mathbf{Z}$*. If* $u = o(n^\tau/\log(p))$*, then*

$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2\right) = o(n^\tau).$$

Proof. Fix $t > 0$ arbitrarily. Define the event $\mathscr{T}_t$ as

$$\mathscr{T}_t \triangleq \bigcap_{\mathbf{m}\in\mathcal{M}_u} \left\{\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2 \leq K_\xi^2\left(u + 2\sqrt{utn^\tau} + 2tn^\tau\right)\right\}.$$

For any fixed $m \in \mathcal{M}_u$, it follows from Lemma 2.12 that

$$\mathbb{P}\left(\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|^2 > K_\xi^2\left(u + 2\sqrt{utn^\tau} + 2tn^\tau\right)\right) \leq \exp\left(-tn^\tau\right).$$

Therefore,

$$\mathbb{P}\left(\mathscr{T}_t^{\mathsf{c}}\right) \leq \exp\left(-tn^\tau + \log\left(|\mathcal{M}_u|\right)\right). \tag{2.6.1}$$

We observe that the above tends to zero from the assumption that $u\log(p) = o(n^\tau)$ and the standard bound on binomial coefficients $|\mathcal{M}_u| = \binom{p}{u} \leq (ep/u)^u$. Now, note that

$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2\right) = \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2 \mathbb{1}_{\mathscr{T}_t}\right) + \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2 \mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\right).$$

For the first term, by the definition of $\mathscr{T}_t$,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2 \mathbb{1}_{\mathscr{T}_t}\right) \leq 2tK_\xi^2.$$

For the second term, by Cauchy-Schwarz and equation (2.6.1), it follows that

$$\begin{aligned}\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\right\|_2^2 \mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\right) &\leq \limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\left\|\boldsymbol{\xi}\right\|_2^2 \mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\right)\\ &\leq \limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\left\|\boldsymbol{\xi}\right\|_2^4\right)^{1/2}\mathbb{P}\left(\mathscr{T}_t^{\mathsf{c}}\right)^{1/2}\\ &= 0.\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\right\|_2^2\right) \leq 2tK_\xi^2.$$

Since $t > 0$ was arbitrary, this finishes the proof. □

Lemma 2.14. *Under the assumptions and setup of Proposition 2.1, for any sub-Gaussian vector $\boldsymbol{\zeta}$ with parameter K_ζ independent of $\mathbf{Z}$,*

1.
$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|_2^2\right) = o(n^\tau).$$

2.
$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right) = o(n^\tau).$$

Note that $\boldsymbol{\zeta}$ is not necessarily independent of $\boldsymbol{\xi}$.

Proof. For $\mathbf{m} \in \mathcal{M}_u$, let

$$r_{\mathbf{m}} \triangleq \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|_2^2.$$

Fixing $t > 0$ arbitrarily, define the set

$$\mathcal{A}_t \triangleq \left\{\mathbf{m} \in \mathcal{M}_u : r_{\mathbf{m}} \leq tn^\tau\right\}.$$

Now,

$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} r_{\mathbf{m}}\right) = \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}} r_{\mathbf{m}}\right) + \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}} w_{\mathbf{m}} r_{\mathbf{m}}\right)$$

By the definition of $\mathcal{A}_t$,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}} r_{\mathbf{m}}\right) \le t.$$

For $\mathcal{A}_t^{\mathsf{c}}$, fix a value of $a > 0$, which will be determined later, and define the set $\mathscr{T}_a$ as

$$\mathscr{T}_a \triangleq \bigcap_{\mathbf{m}\in\mathcal{M}_u} \left\{\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\|_2^2 \le K_\xi^2\left(u + 2\sqrt{uan^\tau} + 2an^\tau\right)\right\}.$$

By the calculations from equation (2.6.1), it follows that

$$\mathbb{P}\left(\mathscr{T}_a^{\mathsf{c}}\right) \le \exp\left(-an^\tau + \log\left(|\mathcal{M}_u|\right)\right). \tag{2.6.2}$$

Moreover, note that, by assumption,

$$\limsup_{n\to\infty}\sup_{\mathbf{m}\in\mathcal{M}_u} n^{-1} r_{\mathbf{m}} \le \limsup_{n\to\infty} n^{-1}\|\boldsymbol{\mu}\|_2^2 \le C,$$

for some constant $C > 0$. Then, for n sufficiently large,

$$n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}} w_{\mathbf{m}} r_{\mathbf{m}}\right) \le 2Cn^{1-\tau}\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\mathbb{E}\left(w_{\mathbf{m}}\right) \le 2Cn^{1-\tau}\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\left(\mathbb{E}\left(w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a}\right) + \mathbb{P}\left(\mathscr{T}_a^{\mathsf{c}}\right)\right). \tag{2.6.3}$$

Fix $\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}$ temporarily and let S be any weakly sparse set for $\boldsymbol{\mu}$. Then, we have that

$$\begin{aligned}
w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a} &\le \exp\left(-\frac{1}{\alpha}\left(\left\|\mathbf{P}_{\mathbf{m}}^{\perp}\mathbf{y}\right\|_2^2 - \left\|\mathbf{P}_{\boldsymbol{S}}^{\perp}\mathbf{y}\right\|_2^2\right)\right)\mathbb{1}_{\mathscr{T}_a} \\
&\le \exp\left(-\frac{1}{\alpha}\left(r_{\mathbf{m}} - r_{\boldsymbol{S}} + 2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\xi} - 2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\boldsymbol{S}}^{\perp}\boldsymbol{\xi} - K_\xi^2\left(u + 2\sqrt{uan^\tau} + 2an^\tau\right)\right)\right).
\end{aligned}$$

By Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\left(w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a}\right) \le &\exp\left(-\frac{1}{\alpha}\left(r_{\mathbf{m}} - r_{\boldsymbol{S}} - K_\xi^2\left(u + 2\sqrt{uan^\tau} + 2an^\tau\right)\right)\right) \\
&\times\left(\mathbb{E}\exp\left(-\frac{4}{\alpha}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\xi}\right)\right)^{1/2}\left(\mathbb{E}\exp\left(\frac{4}{\alpha}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\boldsymbol{S}}^{\perp}\boldsymbol{\xi}\right)\right)^{1/2}.
\end{aligned}$$

Computing each of the Laplace transforms directly, it follows that

$$\mathbb{E}\exp\left(-\frac{4}{\alpha}\boldsymbol{\mu}^\mathsf{T}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\xi}\right) \leq \exp\left(\frac{8K_\xi^2}{\alpha^2}r_{\mathbf{m}}\right).$$

Here, we have used Definition 2.1.2. Similarly,

$$\mathbb{E}\exp\left(\frac{4}{\alpha}\boldsymbol{\mu}^\mathsf{T}\mathbf{P}_{\boldsymbol{S}}^{\perp}\boldsymbol{\xi}\right) \leq \exp\left(\frac{8K_\xi^2}{\alpha^2}r_{\boldsymbol{S}}\right).$$

Hence,

$$\begin{aligned}\mathbb{E}\left(w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a}\right) &\leq \exp\left(-\frac{1}{\alpha}\left(\left(1-\frac{4K_\xi^2}{\alpha}\right)r_{\mathbf{m}} - \left(1+\frac{4K_\xi^2}{\alpha}\right)r_{\boldsymbol{S}} - K_\xi^2\left(u+2\sqrt{uan^\tau}+2an^\tau\right)\right)\right)\\ &\leq \exp\left(-\frac{1}{\alpha}\left(\left(1-\frac{4K_\xi^2}{\alpha}\right)tn^\tau - \left(1+\frac{4K_\xi^2}{\alpha}\right)r_{\boldsymbol{S}} - K_\xi^2\left(u+2\sqrt{uan^\tau}+2an^\tau\right)\right)\right).\end{aligned}$$

The second inequality follows from the fact that $\mathbf{m} \in \mathcal{A}_t^\mathsf{c}$. Since $u = o(n^\tau/\log(p))$, setting $a < \left(1 - 4K_\xi^2/\alpha\right)t/2$ yields

$$\mathbb{E}\left(w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a}\right) \leq \exp\left(-\frac{1}{\alpha}\left(\left(1-\frac{4K_\xi^2}{\alpha}\right)t - 2a\right)n^\tau + o(n^\tau)\right) \tag{2.6.4}$$

Combining equations (2.6.2), (2.6.3), and (2.6.4), it follows that

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t^\mathsf{c}} w_{\mathbf{m}}r_{\mathbf{m}}\right) = 0.$$

Therefore,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}r_{\mathbf{m}}\right) \leq t.$$

Since $t > 0$ was arbitrary, the first claim follows. For the second half, let the set $\mathscr{F}_t$ be

$$\mathscr{F}_t \triangleq \bigcap_{m\in\mathcal{A}_t}\left\{\left|\boldsymbol{\mu}^\mathsf{T}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right| \leq tn^\tau\right\}.$$

For a fixed $\mathbf{m} \in \mathcal{A}_t$, it follows by a Chernoff bound that, for some constant $c > 0$,

$$\mathbb{P}\left(\left|\boldsymbol{\mu}^\mathsf{T}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right| > tn^\tau\right) \leq 2\exp\left(-\frac{ct^2n^{2\tau}}{K_\zeta^2 r_{\mathbf{m}}}\right) \leq 2\exp\left(-\frac{ctn^\tau}{K_\zeta^2}\right).$$

Therefore, an upper bound for $\mathbb{P}\left(\mathscr{F}_t^{\mathsf{c}}\right)$ is given by

$$\mathbb{P}\left(\mathscr{F}_t^{\mathsf{c}}\right) \leq 2\exp\left(-\frac{ctn^{\tau}}{K_{\zeta}^2} + \log(|\mathcal{A}_t|)\right). \tag{2.6.5}$$

Now,

$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\right) = \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\mathbb{1}_{\mathscr{F}_t}\right) + \mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\right).$$

By the definition of $\mathscr{F}_t$, it follows that

$$\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\mathbb{1}_{\mathscr{F}_t}\right) \leq tn^{\tau}.$$

On $\mathscr{F}_t^{\mathsf{c}}$, two applications of Cauchy-Schwarz and equation (2.6.5) yields

$$\begin{aligned}
&\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\right)\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\mathbb{E}\left(\left\|\boldsymbol{\zeta}\right\|_2\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\right)\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\left(\mathbb{E}\left\|\boldsymbol{\zeta}\right\|_2^2\right)^{1/2}\left(\mathbb{P}\left(\mathscr{F}_t^{\mathsf{c}}\right)\right)^{1/2}\\
&\qquad= 0.
\end{aligned}$$

Furthermore, on $\mathcal{A}_t^{\mathsf{c}}$, by another two applications of Cauchy-Schwarz,

$$\begin{aligned}
&\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\left(\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}} w_{\mathbf{m}}\left|\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right|\right)\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\mathbb{E}\left(w_{\mathbf{m}}\left\|\boldsymbol{\zeta}\right\|_2\right)\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\left(\mathbb{E}w_{\mathbf{m}}^2\right)^{1/2}\left(\mathbb{E}\left\|\boldsymbol{\zeta}\right\|_2^2\right)^{1/2}\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\left(\mathbb{E}\left\|\boldsymbol{\zeta}\right\|_2^2\right)^{1/2}\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\left(\mathbb{E}w_{\mathbf{m}}\right)^{1/2}\\
&\qquad\leq \limsup_{n\to\infty} n^{-\tau}\left\|\boldsymbol{\mu}\right\|_2\left(\mathbb{E}\left\|\boldsymbol{\zeta}\right\|_2^2\right)^{1/2}\sum_{\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}}\left(\mathbb{E}\left(w_{\mathbf{m}}\mathbb{1}_{\mathscr{T}_a}\right) + \mathbb{P}\left(\mathscr{T}_a^{\mathsf{c}}\right)\right)^{1/2}\\
&\qquad= 0,
\end{aligned}$$

where the limit follows by equations (2.6.2) and (2.6.4). Since $t > 0$ was arbitrary, this proves the second claim and finishes the proof. $\square$

Immediately, we have the following corollary for random designs when the mean vector is assumed to be weakly sparse in probability. Consider the setup of Lemma 2.14. If $\boldsymbol{\mu}$ is weakly sparse relative to $\mathbf{Z}$ in probability and $\|\boldsymbol{\mu}\|_2^2 = \mathcal{O}_{\mathbb{P}}(n^\tau)$, then

1. $$\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|_2^2\right) = o_{\mathbb{P}}(n^\tau).$$

2. $$\left(\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\zeta}\right) = o_{\mathbb{P}}(n^\tau).$$

With these lemmata, we can now prove Proposition 2.1.

Proof of Proposition 2.1. Indeed, by convexity of the norm, it follows that

$$\left\|\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\mathbf{Z}\hat{\boldsymbol{\gamma}}_{\mathbf{m}} - \boldsymbol{\mu}\right\|_2^2 \leq \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \left\|\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu}\right\|_2^2 + \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}} \left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\xi}\right\|_2^2.$$

Applying Lemmata 2.13 and 2.14 finishes the proof. $\square$

Instead of directly proving Theorem 2.2, we decompose the estimator and prove each part separately. Indeed, we note that

$$\hat{\beta}_{\text{EW}} = \frac{\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}} + \boldsymbol{\eta}\right)^{\mathsf{T}}\left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}} + \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}\right)}{\left\|\mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right\|_2^2}.$$

Then,

$$\begin{aligned}\sqrt{n}\hat{\beta}_{\text{EW}} = \Big(&\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}} + \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}\right) + \boldsymbol{\eta}^{\mathsf{T}}\left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right)\\ &+ \boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\eta}\beta^* + \boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\varepsilon}\Big) \times \frac{1}{\sqrt{n}\sigma_\eta^2} \times \frac{n\sigma_\eta^2}{\left\|\mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right\|^2}.\end{aligned}$$

We start by proving that the first line, which corresponds to the bias from inexact orthogonalization, converges to zero.

Lemma 2.15. *Consider the models given in equations* (2.2.1) *and* (2.2.2). *Under assumptions (2.1) – (2.3),*

$$\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{EW}\right)^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{EW} + \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}\right) + \boldsymbol{\eta}^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{EW}\right) = o_{\mathbb{P}}(\sqrt{n}).$$

Proof. Without the loss of generality, we assume that $u \triangleq u_\theta = u_\delta$. Expanding, we have

$$\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right) + \left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}) + \boldsymbol{\eta}^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right).$$

We consider each of the three terms separately. By Cauchy-Schwarz and Corollary 2.1.1, it follows that

$$\left|\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right)\right| \leq \left\|\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right\|_2 \left\|\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right\|_2 = o_{\mathbb{P}}(\sqrt{n}).$$

For the second term, we may further expand to obtain

$$\begin{aligned}
\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}) &= \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},x} \left(\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\nu} - \mathbf{P}_{\mathbf{m}}\boldsymbol{\eta}\right)^{\mathsf{T}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}) \\
&= \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},x}\boldsymbol{\nu}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}) + \frac{1}{2}\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},x} \left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\varepsilon}\right\|_2^2 \\
&\quad - \frac{1}{2}\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},x} \left\|\mathbf{P}_{\mathbf{m}} (\boldsymbol{\eta} + \boldsymbol{\varepsilon})\right\|_2^2 \\
&\quad - \left(\beta^* - \frac{1}{2}\right)\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},x} \left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\eta}\right\|_2^2.
\end{aligned}$$

In the model $\mathbf{x} = \boldsymbol{\nu} + \boldsymbol{\eta}$, applying Lemma 2.13 with $\boldsymbol{\xi} = \boldsymbol{\varepsilon}$, $\boldsymbol{\xi} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$, and $\boldsymbol{\xi} = \boldsymbol{\eta}$ and Corollary 2.6.1 with $\boldsymbol{\zeta} = \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}$ implies that

$$\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}) = o_{\mathbb{P}}(\sqrt{n}).$$

Finally,

$$\begin{aligned}
\boldsymbol{\eta}^{\mathsf{T}} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right) &= \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},y}\boldsymbol{\eta}^{\mathsf{T}} \left(\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu} - \mathbf{P}_{\mathbf{m}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon})\right) \\
&= \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},y}\boldsymbol{\eta}^{\mathsf{T}}\mathbf{P}_{\mathbf{m}}^{\perp}\boldsymbol{\mu} - \frac{1}{2}\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},y} \left\|\mathbf{P}_{\mathbf{m}} (\boldsymbol{\eta}(\beta^* + 1) + \boldsymbol{\varepsilon})\right\|_2^2 \\
&\quad + \frac{1}{2}\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},y} \left\|\mathbf{P}_{\mathbf{m}} (\boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon})\right\|_2^2 + \frac{1}{2}\sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m},y} \left\|\mathbf{P}_{\mathbf{m}}\boldsymbol{\eta}\right\|_2^2.
\end{aligned}$$

To finish the proof, we similarly apply Lemma 2.13 and Corollary 2.6.1 in the model $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\eta}\beta^* + \boldsymbol{\varepsilon}$. It follows that

$$\boldsymbol{\eta}^\mathsf{T} \left(\boldsymbol{\mu} - \mathbf{Z}\hat{\boldsymbol{\theta}}_{\mathrm{EW}} \right) = o_{\mathbb{P}}(\sqrt{n}).$$

□

Lemma 2.16. *Consider the models given in (2.2.1) and (2.2.2). Under assumptions (2.1)– (2.3),*

1.

$$\sqrt{n} \left(\frac{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\eta}\beta^*}{\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{EW} \right\|_2^2} - \beta^* \right) \xrightarrow{\mathbb{P}} 0.$$

2.

$$n^{-1/2} \frac{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\varepsilon}}{\sigma_\eta^2} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right).$$

3.

$$\frac{n\sigma_\eta^2}{\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{EW} \right\|_2^2} \xrightarrow{\mathbb{P}} 1.$$

Proof. Indeed, expanding the denominator, we see that

$$\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2 = \left\| \boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2 + 2\boldsymbol{\eta}^\mathsf{T}\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right) + \|\boldsymbol{\eta}\|_2^2.$$

By Corollary 2.1.1 and Lemma 2.15, it follows that

$$\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2 = o_{\mathbb{P}}(\sqrt{n}) + \|\boldsymbol{\eta}\|_2^2.$$

Then, by the Law of Large Numbers, $n^{-1}\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2 \xrightarrow{\mathbb{P}} \sigma_\eta^2$. This proves the third claim. Now, by direct substitution, we have that

$$\sqrt{n} \left(\frac{\left(\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2 + o_{\mathbb{P}}(\sqrt{n}) \right)\beta^*}{\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|^2} - \beta^* \right) = \frac{n}{\left\| \mathbf{x} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\mathrm{EW}} \right\|_2^2} \frac{o_{\mathbb{P}}(\sqrt{n})}{\sqrt{n}} = o_{\mathbb{P}}(1),$$

which proves the first claim. The second claim follows by the Central Limit Theorem, which finishes the proof. □

Proof of Theorem 2.2. The proof follows by combining Lemmata 2.15 and 2.16. □

Proof of Corollary 2.2.1. By possibly enlarging $\mathscr{K}$, we note that $\mathscr{K}$ can be written as

$$\mathscr{K} = \{(\beta^*, \sigma_\eta^2, \sigma_\varepsilon^2, K_\eta, K_\varepsilon) : |\beta^*| \le \beta_U^*, \sigma_\eta^2 \in [\sigma_{\eta,L}^2, \sigma_{\eta,U}^2], \sigma_\varepsilon^2 \in [\sigma_{\varepsilon,L}^2, \sigma_{\varepsilon,U}^2], \\ K_\eta \in [K_{\eta,L}, K_{\eta,U}], K_\varepsilon \in [K_{\varepsilon,L}, K_{\varepsilon,U}]\}$$

for fixed positive constants β_U^*, $\sigma_{\eta,L}^2$, $\sigma_{\eta,U}^2$, $\sigma_{\varepsilon,L}^2$, $\sigma_{\varepsilon,U}^2$, $K_{\eta,L}$, $K_{\eta,U}$, $K_{\varepsilon,L}$, and $K_{\varepsilon,U}$. Observe that the vectors $\boldsymbol{\eta}$, $\boldsymbol{\eta}\beta^*$, and $\boldsymbol{\varepsilon}$ are uniformly sub-Gaussian with parameters $K_{\eta,U}$ $\beta_U^* K_{\eta,U}$, and $K_{\varepsilon,U}$ for $\boldsymbol{\vartheta} \in \mathscr{K}$ respectively. Thus, applications of Lemmata 2.13 and 2.14 are uniform. Therefore, Lemmata 2.15 and 2.16 also hold uniformly for $\boldsymbol{\vartheta} \in \mathscr{K}$, which will prove the claim. □

2.6.2 Proofs for Section 2.2.4

Proof of Theorem 2.6. Suppose that $s_\delta = o(\sqrt{n}/\log(p))$. We consider a sequence of $\in \boldsymbol{\Theta}(s_\gamma, s_\delta)$ such that $\boldsymbol{S}_\gamma \cap \boldsymbol{S}_\delta = \varnothing$ and $\boldsymbol{\delta}^* \ge \mathbf{0}_n$ componentwise. We construct $\boldsymbol{\Sigma}_{Z,Z}$ implicitly. For $j \in \boldsymbol{S}_\delta^{\mathsf{c}}$, let

$$\mathbf{Z}_j \overset{i.i.d.}{\sim} \mathcal{N}_n\left(\mathbf{0}_n, \mathbf{I}_n\right).$$

Before defining $\mathbf{Z}_j$ for $j \in \boldsymbol{S}_\delta$, we need to define another Gaussian matrix $\Xi \in \mathbb{R}^{n\times p}$. For $j \in \boldsymbol{S}_\delta^{\mathsf{c}}$, set $\Xi_j = 0_n$. Then, for $j \in \boldsymbol{S}_\delta$,

$$\Xi_j \overset{i.i.d.}{\sim} \mathcal{N}_n\left(\mathbf{0}_n, \tau_n^2 \mathbf{I}_n\right),$$

independent of $\mathbf{Z}_k$ for all $k \in \boldsymbol{S}_\delta^{\mathsf{c}}$; the value $\tau_n^2 > 0$ is determined later. Now, for $j \in \boldsymbol{S}_\delta$, we let $\mathbf{Z}_j = \mathbf{Z}\boldsymbol{\gamma}^* + \Xi_j$. Therefore, it follows that $\mathbf{Z}\boldsymbol{\delta}^* = \mathbf{Z}\boldsymbol{\gamma}^* \left\|\boldsymbol{\delta}^*\right\|_1 + \Xi\boldsymbol{\delta}^*$. By a direct calculation,

$$\mathrm{Cov}((\mathbf{Z}\boldsymbol{\delta}^*)_1, (\mathbf{Z}\boldsymbol{\gamma}^*)_1) = \mathrm{Cov}((\mathbf{Z}\boldsymbol{\gamma}^*)_1 \left\|\boldsymbol{\delta}^*\right\|_1 + (\Xi\boldsymbol{\delta}^*)_1, (\mathbf{Z}\boldsymbol{\gamma}^*)_1) = \mathrm{Var}((\mathbf{Z}\boldsymbol{\gamma}^*)_1) \left\|\boldsymbol{\delta}^*\right\|_1.$$

Moreover,

$$\mathrm{Var}((\mathbf{Z}\boldsymbol{\delta}^*)_1) = \mathrm{Var}((\mathbf{Z}\boldsymbol{\gamma}^*)_1 \left\|\boldsymbol{\delta}^*\right\|_1 + (\Xi\boldsymbol{\delta}^*)_1) = \mathrm{Var}((\mathbf{Z}\boldsymbol{\gamma}^*)_1) \left\|\boldsymbol{\delta}^*\right\|_1^2 + \tau_n^2 \left\|\boldsymbol{\delta}^*\right\|_2^2.$$

Choosing $\tau_n^2 \to 0$ sufficiently fast, it follows that

$$\mathrm{Var}((\mathbf{Z}\boldsymbol{\delta}^*)_1) = \mathrm{Var}((\mathbf{Z}\boldsymbol{\gamma}^*)_1) \left\|\boldsymbol{\delta}^*\right\|_1^2 + o\left(n^{-1/2}\right).$$

Hence, this implies that

$$\mathrm{Cov}\left(\left(\mathbf{Z}\boldsymbol{\delta}^*\right)_1, \left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1\right) = \sqrt{\mathrm{Var}(\left(\mathbf{Z}\boldsymbol{\delta}^*\right)_1)\mathrm{Var}(\left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1)} + o(n^{-1/2}).$$

Now, note that

$$\mathrm{Cov}\left(\left(\mathbf{Z}\boldsymbol{\delta}^*\right)_1, \left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1\right) = \mathrm{Cov}\left(\mathbf{x}_1, \mathbf{y}_1\right) - \beta^*\mathrm{Var}\left(\mathbf{x}_1\right).$$

Let $\hat{\beta}$ be any $\sqrt{n}$-consistent estimator for β. Then, $n^{-1}\left(\mathbf{x}^\mathsf{T}\mathbf{y} - \hat{\beta}\mathbf{x}^\mathsf{T}\mathbf{x}\right)$ is a $\sqrt{n}$-consistent estimator for $\mathrm{Cov}\left(\left(\mathbf{Z}\boldsymbol{\delta}^*\right)_1, \left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1\right)$. Consider an oracle that has access to the set $\boldsymbol{S}_\delta$, knows $\boldsymbol{S}_\delta \cap \boldsymbol{S}_\gamma = \varnothing$, and knows the covariance structure of the design. Then, since $s_\delta = o(\sqrt{n}/\log(p))$, a $\sqrt{n}$-consistent estimator for $\mathrm{Var}(\left(\mathbf{Z}\boldsymbol{\delta}^*\right)_1)$ is given by Theorem 2.8. This implies that there exists a $\sqrt{n}$-consistent estimator for $\mathrm{Var}(\left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1)$. By the minimax lower bounds established by Cai and Guo (2020), it follows that, in order to have a $\sqrt{n}$-consistent estimator for $\mathrm{Var}(\left(\mathbf{Z}\boldsymbol{\gamma}^*\right)_1)$, it must be the case that $s_\gamma = \mathcal{O}(\sqrt{n}/\log(p))$. This proves half of the claim. The other half follows by symmetry, which finishes the proof. $\square$

CHAPTER 3

Inference and Estimation for Random Effects in High-Dimensional Linear Mixed Models

3.1 Introduction

In the past two decades, there has been a lot of progress in the theory for high-dimensional linear models. However, its close cousin, the high-dimensional linear mixed model, has received significantly less attention; it was not until the past decade until there were procedures for estimation. Consider a linear mixed model given by

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{3.1.1}$$

with $\mathbf{Z} \in \mathbb{R}^{n\times q}$, $\mathbf{W} \in \mathbb{R}^{n\times d}$, and $\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$; the vector $\boldsymbol{\mu}$ and the pair $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ are the fixed effects and the random effects respectively. In addition, we observe covariates $\mathbf{X} \in \mathbb{R}^{n\times p}$ such that $\boldsymbol{\mu} \approx \mathbf{X}\boldsymbol{\beta}^*$ for some sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ (see Section 3.1.2 for a rigorous definition). Here, $\mathbf{X}$ is the component of the design corresponding to the fixed effects and $(\mathbf{Z}, \mathbf{W})$ the component corresponding to the random effects. We consider the setting where the random effects are low-dimensional, $q + d < n$, but the fixed effects are high-dimensional, $p > n$. We have separated the random effects in two to emphasize that later we are interested in $\boldsymbol{\nu}$ and view $\boldsymbol{\gamma}$ as nuisance parameters. Various authors have considered different aspects of this problem.

The earliest work of Schelldorfer et al. (2011) proposed an estimator for both $\boldsymbol{\beta}^*$ and the variance components using a lasso-type approach. These types of approaches were later extended by several authors who considered estimation with both convex penalties, such as Groll and Tutz (2014), and non-convex penalties, such as Wang et al. (2012). There is also a growing literature on model selection in high-dimensional linear mixed models (for example, see the review article by Müller et al. (2013)).

The problem of inference is slightly less well studied. To the best of our knowledge, hypotheses testing problems were first considered by Chen et al. (2015) for random effects and Bradic et al. (2019) for fixed effects. However, the work of Chen et al. (2015) only consider the special case of ANOVA designs for random effects. During the preparation of this manuscript, we became aware of the independent work of Li et al. (2019), who consider the problem of inference in high-dimensional linear mixed models. In particular, they discuss inference for fixed effects and estimation of variance components. A more detailed comparison of our methodology with Li et al. (2019) is deferred to Section 3.2.4. We also note that there is a parallel notion of high-dimensional mixed models, where the number of fixed effects is low-dimensional while the random effects are high-dimensional. Under this setting, Jiang et al. (2016) established asymptotic results for the restricted maximum likelihood for variance components.

The goal of the present paper is to contribute to this growing literature on high-dimensional linear mixed models where the fixed effects are high-dimensional, both in terms of estimation and inference. In particular, we consider three related problems:

1. Testing whether a collection of random effects is zero.

2. Constructing confidence intervals for the variance of a single random effect.

3. Estimating using empirical Bayes in Gaussian ANOVA Type Models.

Our methodology is inspired by both low-dimensional linear mixed models as well as high-dimensional linear models. Specifically, our approach to all three problems starts with considering a procedure in the corresponding low-dimensional problem and retrofitting it with tools and techniques from high-dimensional linear models to produce a procedure for high-dimensional linear mixed models. Throughout the paper, while we consider the general linear mixed effects models, we use the balanced one-way ANOVA model to simplify the discussion of our estimators and assumptions.

3.1.1 Organization of the Chapter

We end the current section with a description of the notation that we adopt throughout the paper. Sections 3.2, 3.3, and 3.4 consider the three problems outlined in the Introduction in succession. Each one starts with a description of the problem setup, a brief motivation from the low-dimensional problem, and a description of the estimator that is considered, and ends with some theoretical results. In Sections 3.5 and 3.6, we provide the results of our simulations and a real data application respectively. For the ease of presentation, we defer all proofs and additional simulation results to Appendix 2.

3.1.2 Notation

Throughout, all of our variables have a dependence on n, but we suppress this dependence when it does not cause confusion. For a general vector $\mathbf{a}$ and matrix $\mathbf{A}$, let $\|\mathbf{a}\|_2$ denote the standard Euclidean norm with the dimension of the space being implicit from the vector, $\|\mathbf{A}\|_2$ the operator norm, and $\|\mathbf{A}\|_{\text{HS}}$ the Hilbert-Schmidt norm. Furthermore, if $\mathbf{A}$ is square, then $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximal and minimal eigenvalue of $\mathbf{A}$ respectively. For any $k \in \mathbb{N}$, we let $\lambda_{\max,k}(\mathbf{A})$ denote the kth largest eigenvalue of $\mathbf{A}$ if $\mathbf{A}$ is square. Moreover, we write $\mathbf{1}_k \in \mathbb{R}^k$ and $\mathbf{I}_k \in \mathbb{R}^{k\times k}$ to denote the k-dimensional vector of all ones and the k-dimensional identity matrix respectively. For two matrices $\mathbf{A}$ and $\mathbf{B}$, the notation $\mathbf{A} \ominus \mathbf{B}$ denotes the intersection of the column space of $\mathbf{A}$ and the orthogonal complement of the column space of $\mathbf{B}$. Then, for a matrix $\mathbf{A}$, we write $\mathbf{P}_{\mathbf{A}}$ to denote the projection onto the column space of $\mathbf{A}$ and $\mathbf{P}_{\mathbf{A}}^{\perp}$ the projection onto the orthogonal complement. Moreover, we write $r_{\mathbf{A}}$ to denote the rank of $\mathbf{A}$.

Consistent with other high-dimensional works, we assume that $\boldsymbol{\beta}^*$ is a sparse vector. There are various notions of sparsity, but we assume the general setting of weak sparsity from Chapter 2. For $u \in \mathbb{N}$, we let $\mathcal{M}_u \triangleq \{\mathbf{m} \subseteq \{1,\dots,p\} : |\mathbf{m}| = u\}$ denote the collection of all models with the dimension of the fixed effects design equal to u. For a model $\mathbf{m} \in \mathcal{M}_u$, $\mathbf{X}_{\mathbf{m}}$ denotes the $n \times u$ sub-matrix of $\mathbf{X}$ corresponding to the columns indexed by $\mathbf{m}$. Then we write $\mathcal{S}_\mu \triangleq \left\{\mathbf{m} \in \mathcal{M}_s^* : \|\mathbf{P}_{\mathbf{X}_{\mathbf{m}}}^{\perp}\boldsymbol{\mu}\|_2^2 = o(k)\right\}$ and let $\boldsymbol{S} \in \mathcal{S}_\mu$ denote any weakly sparse set for $\boldsymbol{\mu}$. We note that the usual high-dimensional setting of strong sparsity, where $\boldsymbol{\mu} = \mathbf{X}_{\boldsymbol{S}}\boldsymbol{\beta}_{\boldsymbol{S}}^*$ for $|\boldsymbol{S}| = s^*$, implies that $\boldsymbol{\mu}$ is weakly sparse relative to $\mathbf{X}$ with sparsity s^*.

Note that if $\boldsymbol{\xi}$ is sub-Gaussian with parameter K and $\mathbf{A} \in \mathbb{R}^{a\times n}$ is any deterministic matrix, then $\mathbf{A}\boldsymbol{\xi}$ is also sub-Gaussian with parameter $K\,\lambda_{\max}(\mathbf{A}^\top\mathbf{A})$. Finally, the asymptotic distributions of some of our estimators depend on the fourth moments of the underlying distributions. We write $\kappa_\varepsilon \triangleq \text{Var}(\boldsymbol{\varepsilon}_1^2)$, $\omega_\varepsilon \triangleq \mathbb{E}(\boldsymbol{\varepsilon}_1^4)$, $\kappa_\nu \triangleq \text{Var}(\boldsymbol{\nu}_1^2)$, and $\omega_\nu \triangleq \mathbb{E}(\boldsymbol{\nu}_1^4)$ when $\boldsymbol{\nu}$ corresponds to a single random effect.

3.2 Hypotheses Testing for Random Effects

In this section, we consider the problem of inference for a collection of random effects. Consider the high-dimensional linear mixed model (3.1.1) and let $\boldsymbol{\Psi} \triangleq \text{Var}(\boldsymbol{\nu})$. We are interested in the hypotheses testing problem

$$H_0 : \lambda_{\max}(\boldsymbol{\Psi}) = 0, \qquad\qquad H_1 : \lambda_{\max}(\boldsymbol{\Psi}) > 0. \tag{3.2.1}$$

We propose two procedures in this section depending on whether $\boldsymbol{\varepsilon}$ is Gaussian.

3.2.1 Model and Motivation

Suppose temporarily that we are in the low-dimensional Gaussian setting with $s^* = p, p+q+d < n$, $\boldsymbol{\mu} = \mathbf{X}_{\boldsymbol{S}}\boldsymbol{\beta}^*_{\boldsymbol{S}}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2_\varepsilon \mathbf{I}_n)$ for some positive constant $\sigma^2_\varepsilon > 0$, and $\boldsymbol{\nu} \sim \mathcal{N}_q(\mathbf{0}_q, \boldsymbol{\Psi})$ for some symmetric positive semi-definite matrix $\boldsymbol{\Psi}$. Then, in this problem, the standard procedure for testing $\boldsymbol{\nu}$ is through the Wald F-test. Writing $r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})} \triangleq \operatorname{rank}(\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})})$ and $r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp} \triangleq \operatorname{rank}(\mathbf{P}^\perp_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})})$, the Wald F-test is defined as

$$F_{\text{ld}} = \frac{\|\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}\mathbf{y}\|_2^2 / r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}}{\|\mathbf{P}^\perp_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})}\mathbf{y}\|_2^2 / r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}}. \tag{3.2.2}$$

Under the null hypothesis, the above statistic has an $F_{r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}, r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}}$ distribution. The main obstacle to directly using the Wald F-test in the high-dimensional setting is removing the contribution of the fixed effects. One possibility is to perform model selection and choose the relevant covariates from $\mathbf{X}$ and then use the Wald F-test. Chen et al. (2015) consider a similar problem in the growing dimensional setting and they use a SCAD based approach for variable selection. As a consequence, they require $p = o(\sqrt{n})$. Instead, we leverage the fact that a projection onto a particular space is a regression onto a design whose columns span the same space.

Expanding both the numerator and the denominator of the Wald F-statistic, we have that

$$\begin{aligned}
\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}\mathbf{y} &= \mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}\mathbf{Z}\boldsymbol{\nu} + P_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}\boldsymbol{\varepsilon}, \\
\mathbf{P}^\perp_{(\mathbf{X},\mathbf{Z},\mathbf{W})}\mathbf{y} &= \mathbf{P}^\perp_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}.
\end{aligned}$$

In both matrices above, they project onto the orthogonal complement of $\mathbf{W}$, which may still be achieved in the high-dimensional problem since $\mathbf{W}$ is a low-dimensional matrix. Thus, we may find two projection matrices, $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}$ and $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}$, such that

$$\begin{aligned}
\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y} &= \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^* + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\nu} + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\boldsymbol{\varepsilon}, \\
\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{y} &= \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\boldsymbol{\beta}^* + \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}.
\end{aligned}$$

If $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}$ was low-dimensional, obtaining the projection of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y}$ onto the orthogonal complement of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}$ is equivalent to finding the residuals of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y}$ using the covariates $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}$; this yields $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y} - \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}^*$ is the least-squares estimator for $\boldsymbol{\beta}^*$. The same holds for $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}$ and $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{y}$. Then, we have that

$$\begin{aligned}
\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y} - \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}^* &= (\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^* - \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}^*) + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\nu} + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\boldsymbol{\varepsilon}, \\
\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{y} - \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\hat{\boldsymbol{\beta}}^* &= (\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\boldsymbol{\beta}^* - \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\hat{\boldsymbol{\beta}}^*) + \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}.
\end{aligned}$$

Hence, this recasts the problem into one of high-dimensional prediction, for which there have been many procedures suggested to estimate $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^*$ and $\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\boldsymbol{\beta}^*$, such as the lasso and exponential weighting (cf. Tibshirani (1996) and Leung and Barron (2006) respectively). Therefore, we propose using a plug-in estimator for $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^*$ and $\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\boldsymbol{\beta}^*$ using exponential weighting of all models of a particular size and then consider the resultant residuals. Since we view the fixed effects as nuisance parameters, we consider exponential weighting instead of the lasso since exponential weighting does not require any assumptions on the design matrix $\mathbf{X}$. However, most of the theory developed also applies to other plug-in estimators, albeit with simple modifications and much stronger conditions. This idea, under some mild assumptions, provides an asymptotic F-test.

However, there are two asymptotic regimes for the random effects: (i) the number of random effects increases to infinity and (ii) the number of random effects stays bounded. These two settings require slightly different analyses, so we consider separate exponential weighting estimators for the two cases.

Besides providing an asymptotic F distribution when $\boldsymbol{\varepsilon}$ is Gaussian, the F-ratio in equation (3.2.2) simultaneously removes the scaling effect from σ^2_{ε}. When $\boldsymbol{\varepsilon}$ is known only to be sub-Gaussian, the ratio no longer follows an F-distribution. However, after appropriate rescaling, we may still achieve the ancillary property relative to σ^2_{ε} by looking at the difference instead of the ratio. This approach, under slightly stronger sparsity assumptions, leads to an asymptotic z-test with only the sub-Gaussian assumption on the error distribution.

3.2.2 Estimator

In the setting where the number of random effects increases to infinity, instead of estimating $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^*$ and $\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}\mathbf{X}\boldsymbol{\beta}^*$ separately, we estimate $\mathbf{P}^{\perp}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^*$ and then project the resultant vector onto $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}$ and $\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}$ respectively. In addition to saving on computational time by only using exponential weighting once, this also allows us to leverage a larger sample size when estimating the mean vector. To apply exponential weighting, we fix a sequence of sparsities $u = u_n$. Let $\hat{\boldsymbol{\beta}}_{\mathbf{m}}$ denote the least-squares estimator of $\boldsymbol{\beta}^*$ using the model $\mathbf{m} \in \mathcal{M}_u$ with covariates $\mathbf{P}^{\perp}_{\mathbf{W}}\mathbf{X}_{\mathbf{m}}$. Let $K_{\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}}$ denote the sub-Gaussian parameter for $\mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon}$. We define the exponential weights by

$$w_{\mathbf{m}} \triangleq \frac{\exp\left(-\frac{1}{\alpha}\|\mathbf{P}^{\perp}_{\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{m}})\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_u}\exp\left(-\frac{1}{\alpha}\|\mathbf{P}^{\perp}_{\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{k}})\|_2^2\right)},$$

where $\alpha > 4K_{\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}}$. Then, the estimator for $\boldsymbol{\beta}^*$ is given by

$$\hat{\boldsymbol{\beta}}_{\text{EW}} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_u} w_{\mathbf{m}}\hat{\boldsymbol{\beta}}_{\mathbf{m}}.$$

Note that the bound on α is to ensure $\mathbf{P}_{\mathbf{W}}^{\perp}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}}$ is a consistent estimator of $\mathbf{P}_{\mathbf{W}}^{\perp}\mathbf{X}\boldsymbol{\beta}^*$. In the case where both $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ are Gaussian, the above bound on α becomes $\alpha > 4\left(\sigma_{\varepsilon}^2 + \lambda_{\max}\left(\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^{\mathsf{T}}\right)\right)$. Then, we estimate $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\boldsymbol{\beta}^*$ and $\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\mathbf{X}\boldsymbol{\beta}^*$ by $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}}$ and $\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}}$ respectively. The corresponding F-statistic is

$$F_{\text{EW}} \triangleq \frac{\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2/r_{\mathbf{Z}\ominus\mathbf{W}}}{\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2/r_{(\mathbf{Z},\mathbf{W})^{\perp}}}.$$

Similar to the Wald F-statistic, we reject the null hypothesis for large values of F_{EW}. In particular, for a value $\delta \in (0,1)$, let $F_{a,b,\delta}$ denote the δ upper quantile of the $F_{a,b}$ distribution. Then, we consider tests of the form

$$\varphi_{F,\delta} \triangleq \mathbb{1}\left(F_{\text{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^{\perp}},\delta}\right).$$

For the second setting where the number of random effects stay bounded, we estimate the numerator differently. Let $\mathbf{U}_{(\mathbf{Z},\mathbf{W})^{\perp}} \in \mathbb{R}^{n\times r_{(\mathbf{Z},\mathbf{W})^{\perp}}}$ be any orthogonal matrix such that $\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp} = \mathbf{U}_{(\mathbf{Z},\mathbf{W})^{\perp}}\mathbf{U}_{(\mathbf{Z},\mathbf{W})^{\perp}}^{\mathsf{T}}$; for example, the matrix $\mathbf{U}_{(\mathbf{Z},\mathbf{W})^{\perp}}$ may be computed by taking the spectral decomposition of $\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}$. Define $\tilde{\mathbf{y}} = \mathbf{U}_{(\mathbf{Z},\mathbf{W})^{\perp}}^{\mathsf{T}}\mathbf{y}$ and let $\tilde{\mathbf{y}}^{(1)}, \tilde{\mathbf{y}}^{(2)} \in \mathbb{R}^{r_{(\mathbf{Z},\mathbf{W})^{\perp}}/2}$ be a partition of $\tilde{\mathbf{y}}$. We similarly define $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$. Then, letting $\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)}$ (respectively $\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(2)}$) denote the least-squares estimator of $\boldsymbol{\beta}^*$ using the model $\mathbf{m} \in \mathcal{M}_u$ with covariates $\tilde{\mathbf{X}}_{\mathbf{m}}^{(1)}$ (respectively $\tilde{\mathbf{X}}_{\mathbf{m}}^{(2)}$), the exponential weights are defined as

$$\tilde{w}_{\mathbf{m}}^{(1)} \triangleq \frac{\exp\left(-\frac{1}{\tilde{\alpha}}\|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbf{m}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)}\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_u}\exp\left(-\frac{1}{\tilde{\alpha}}\|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbf{k}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{k}}^{(1)}\|_2^2\right)}$$

and similarly for $\tilde{w}_{\mathbf{m}}^{(2)}$, where $\tilde{\alpha}$ is delineated in Theorem 3.4 below. Now, define

$$\tilde{\boldsymbol{\beta}}_{\text{EW}}^{(1)} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_u}\tilde{w}_{\mathbf{m}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(2)}, \qquad \tilde{\boldsymbol{\beta}}_{\text{EW}}^{(2)} \triangleq \sum_{\mathbf{m}\in\mathcal{M}_u}\tilde{w}_{\mathbf{m}}^{(2)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)}.$$

Then, the estimator of $\boldsymbol{\beta}^*$ is

$$\tilde{\boldsymbol{\beta}}_{\text{EW}} \triangleq (\tilde{\boldsymbol{\beta}}_{\text{EW}}^{(1)} + \tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(2)})/2$$

and the corresponding F-statistic is

$$\tilde{F}_{\text{EW}} \triangleq \frac{\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}_{\text{EW}})\|_2^2/r_{\mathbf{Z}\ominus\mathbf{W}}}{\|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2/r_{(\mathbf{Z},\mathbf{W})^{\perp}}}.$$

At first sight, computation of these estimators may seem prohibitive since we need to aggregate over $\binom{p}{u}$ models. However, they may be well approximated by Algorithm 1 from Chapter 2.

In the setting where the $\boldsymbol{\varepsilon}$ is not distributed Gaussian, we consider the following z-statistic

$$z_{\text{EW}} \triangleq \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 - r_{\mathbf{Z}\ominus\mathbf{W}}r^{-1}_{(\mathbf{Z},\mathbf{W})^{\perp}}\|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2.$$

Under proper scaling, the statistic z_{EW} has an asymptotic Gaussian distribution under the null hypothesis. Let

$$\sigma^2_{\varsigma,z} \triangleq \kappa_\varepsilon\sum_{i=1}^n(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} - r_{\mathbf{Z}\ominus\mathbf{W}}r^{-1}_{(\mathbf{Z},\mathbf{W})^{\perp}}\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})})^2_{i,i} + 2\sigma_\varepsilon^4\sum_{i\neq j}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} - r_{\mathbf{Z}\ominus\mathbf{W}}r^{-1}_{(\mathbf{Z},\mathbf{W})^{\perp}}\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})})^2_{i,j},$$

with $\hat{\sigma}^2_{\varsigma,z}$ a consistent estimator of $\sigma^2_{\varsigma,z}$. The quantity $\sigma^2_{\varsigma,z}$ is the scaling factor to ensure a central limit for z_{EW}. Then, letting z_δ denote the δ upper quantile of the standard Gaussian distribution, we consider tests of the form

$$\varphi_{z,\delta} \triangleq \mathbb{1}\left(z_{\text{EW}} > z_\delta\hat{\sigma}_{\varsigma,z}\right).$$

A general discussion regarding $\hat{\sigma}^2_{\varsigma,z}$ is deferred to Section 3.3.4. When ε is not Gaussian, we only consider the setting where the number of random effects increases to infinity since the analysis of z_{EW} relies of a central limit theorem for quadratic forms.

As mentioned in Section 3.2.1, under appropriate conditions, we may also use the lasso instead of exponential weighting. For a suitable choice of $\lambda > 0$, define the lasso estimator of β^* as

$$\hat{\boldsymbol{\beta}}_{\text{LA}} \triangleq \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\mathbf{P}^{\perp}_{\mathbf{W}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Then, the corresponding F-statistic is

$$F_{\text{LA}} \triangleq \frac{\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LA}})\|_2^2/r_{\mathbf{Z}\ominus\mathbf{W}}}{\|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LA}})\|_2^2/r_{(\mathbf{Z},\mathbf{W})^{\perp}}}.$$

3.2.3 Assumptions

In this section, we make the following assumptions.

(3.1) The mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_n$ has squared norm, $\|\boldsymbol{\mu}_n\|_2^2/n$, that is bounded.

(3.2) The vector $\boldsymbol{\varepsilon}$ is sub-Gaussian with parameter K_ε and has independent components.

(3.2*) The vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_n\right)$.

(3.3) The random effects $\boldsymbol{\nu}$ are sub-Gaussian with parameter K_ν.

(3.3*) The random effects $\boldsymbol{\nu}$ satisfy $\mathbf{Z}\boldsymbol{\nu} \sim \mathcal{N}_n\left(\mathbf{0}_n, \mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^\mathsf{T}\right)$.

(3.4) The matrix $\mathbf{Z}$ satisfies $\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T})$ being bounded and $(\mathbf{Z}, \mathbf{W})$ is independent of $\mathbf{X}$.

(3.5) The mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_n$ is weakly sparse relative to $\mathbf{X}$ with sparsity s_n^*, with weak sparsity defined in Definition 2.1.1. Furthermore, the statistician chooses a sequence of sparsities u_n such that $u_n \geq s_n^*$ for n sufficiently large and $u_n = o\left(n^\tau / \log(p)\right)$ for some $\tau \in [1/2, 1]$. Moreover, the number of observations in the reduced models, $r_{\mathbf{Z}\ominus\mathbf{W}}$ and $r_{(\mathbf{Z},\mathbf{W})^\perp}$, satisfy $r_{\mathbf{Z}\ominus\mathbf{W}} \asymp r_{(\mathbf{Z},\mathbf{W})^\perp} \asymp n$.

(3.6) The mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_n$ satisfies $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}^*$ with $\|\boldsymbol{\beta}^*\|_0 = s_n^*$. Furthermore, the statistician chooses a sequence of sparsities u_n such that $u_n \geq s_n^*$ for n sufficiently large and $u_n = o\left(n^\tau / \log(p)\right)$ for some $\tau \in [1/2, 1]$. Moreover, the rows of $\mathbf{X}$ are independent and identically distributed $\mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$ with $\max(\operatorname{diag}(\boldsymbol{\Sigma}_X)) = \mathcal{O}(1)$ and $r_{(\mathbf{Z},\mathbf{W})^\perp} \asymp n$.

(3.7) The vector $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_n$ satisfies

$$\inf_n \left\{ \left(\min_{i=1,\ldots,n} \operatorname{Var}(\boldsymbol{\varepsilon}_{n,i}) \right) \wedge \left(\min_{i=1,\ldots,n} \operatorname{Var}(\boldsymbol{\varepsilon}_{n,i}^2) \right) \right\} > 0,$$
$$\lim_{x\to\infty} \sup_n \left\{ \left(\max_{i=1,\ldots,n} \mathbb{E}\left(\boldsymbol{\varepsilon}_{n,i}^2 : |\boldsymbol{\varepsilon}_{n,i}| > x\right) \right) \vee \left(\max_{i=1,\ldots,n} \mathbb{E}\left(\boldsymbol{\varepsilon}_{n,i}^4 : |\boldsymbol{\varepsilon}_{n,i}| > x\right) \right) \right\} = 0.$$

Remark. Assumptions (3.1) and (3.2) are standard assumptions in high-dimensional linear models. Calling $\boldsymbol{\varepsilon}$ the noise and $\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}$ the random component, assumption (3.1) is a scaling assumption to ensure the ratio of the fixed components to the random components remains bounded asymptotically and is analogous to the assumptions of Bradic et al. (2019), who assume that the population covariance matrix of $\mathbf{X}$ has bounded maximal eigenvalue and $\|\boldsymbol{\beta}^*\|_2 = \mathcal{O}(1)$. Next, (3.2) is used for consistency of the prediction procedure under the null hypothesis and assumption (3.3) allows for concentration of the prediction procedure under the alternative hypothesis. Both assumptions are used by various authors, such as Bradic et al. (2019) and Cai and Guo (2017). We note that (3.2) and (3.3) are implied by (3.2*) and (3.3*) respectively, but the additional Gaussian distribution assumption allows us to relate our methodology to the vast literature on low-dimensional Gaussian mixed models.

Next, the first part of assumption (3.4), like (3.1), ensures that the ratio of the fixed components to the random components of the variance noise ratio remains bounded under the alternative hypothesis. To elucidate this point, consider the Gaussian setting with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma_\varepsilon^2 I_n)$ and $\boldsymbol{\nu} \sim \mathcal{N}_q(\mathbf{0}_q, \sigma_\nu^2 I_q)$. Then, $\mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma_\nu^2 Z Z^\mathsf{T} + \sigma_\varepsilon^2 I_n)$. Since $\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T}) \geq \max(\mathrm{diag}(\mathbf{Z}\mathbf{Z}^\mathsf{T}))$, assumption (3.4) bounds the variance of the noise. Moreover, (3.3) and (3.4) together imply that $\mathbf{Z}\boldsymbol{\nu}$ is sub-Gaussian with parameter $K_\nu \, \lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T})$. This requirement is similar to Condition 1 of Bradic et al. (2019) and Condition 3.1 of Cai and Guo (2017). The assumption that $(\mathbf{Z}, \mathbf{W})$ is independent of $\mathbf{X}$ is common in the literature (see the discussion before Condition 3.2 of Cai and Guo (2017)).

The following two assumptions, (3.5) and (3.6), are about the sparsity of the fixed effects. The two assumptions consider different asymptotic regimes regarding the random effects; (3.5) assumes that the number of random effects increases to infinity while (3.6) allows for the number of random effects to stay bounded. The first part of both (3.5) and (3.6) is a sparsity assumption commonly found in the high-dimensional linear models literature, which is discussed further in Remark 3.2.4 below. Note that since the selected sequence of sparsities u_n satisfies $u_n = o(n^\tau / \log(p))$, then the true sequence of sparsities s_n^* also satisfies the same requirement.

The second half of (3.5) is an assumption on the component of the design for the random effects, requiring the number of realizations of the random effects to increase to infinity. The requirement that $r_{\mathbf{Z}\ominus\mathbf{W}} \asymp r_{(\mathbf{Z},\mathbf{W})^\perp} \asymp n$ is for convenience and can be weakened to only $\min(r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^\perp}) \to \infty$ if the sparsity requirement is accordingly relaxed to $u_n = o\big(\min(r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^\perp})^\tau / \log(p)\big)$. The second half of (3.6) is a technical requirement to ensure consistency of exponential aggregation for out-of-sample predictions. Since the number of random effects remains bounded, the regression of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{y}$ on $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}$ does not necessarily yield a consistent estimator of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\boldsymbol{\mu}$ for arbitrary designs. With the Gaussian assumption, we may estimate $\boldsymbol{\beta}^*$ by regressing $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{y}$ on $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\mathbf{X}$ and obtain a consistent estimator of $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\boldsymbol{\mu}$. Again, the requirement that $r_{(\mathbf{Z},\mathbf{W})^\perp} \asymp n$ can be weakened to $r_{(\mathbf{Z},\mathbf{W})^\perp} \to \infty$ by adjusting the sparsity requirement to $u_n = o(r^\tau_{(\mathbf{Z},\mathbf{W})^\perp} / \log(p))$.

Assumption (3.7) is a mild assumption on the distribution of $\boldsymbol{\varepsilon}$ to ensure a central limit theorem. For example, (3.7) is satisfied by the Gaussian distribution. Note that no assumption is necessary on $\boldsymbol{\gamma}$ as the nuisance parameters are projected out in the first stage.

Example 1 (Balanced one-way ANOVA)**.** As an example of a design satisfying the above assumptions on , consider a balanced one-way ANOVA design, with q subjects, m observations per subject, and $n = mq$ total observations. In this setting, there are no nuisance random effects, so $d = 0$. Assume further that the number of observations per subject remains bounded (ie., $m = \mathcal{O}(1)$), which is commonly satisfied in practice. Then, the matrix $\mathbf{Z}$ may be represented by $\mathbf{Z} = \mathbf{I}_q \otimes \mathbf{1}_m$. It is immediate that $r_{\mathbf{Z}\ominus\mathbf{W}} = q$ and $r_{(\mathbf{Z},\mathbf{W})^\perp} = (m-1)q$, implying that the second half of (3.5) is satisfied. Finally, assumption (3.4) is satisfied since $\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T}) = \lambda_{\max}(m\mathbf{I}_q) = m$.

3.2.4 Main Results

Since F_{EW} is motivated by the classical F-statistic F_{ld}, the following theorem shows that, up to a small bias term depending on the sparsity, the two statistics are asymptotically equivalent.

Theorem 3.1. *Consider the model given in equation (3.1.1) and the hypotheses testing problem from equation (3.2.1). Assume (3.1), (3.2*), (3.3*), (3.4), and (3.5). If* $\alpha \geq 4\left(\sigma_\varepsilon^2 + \lambda_{max}\left(\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^\intercal\right)\right)$, *then*

$$F_{EW} = F_{ld} + o_{\mathbb{P}}(n^{\tau-1}).$$

As mentioned in Section 3.2.1, under the null hypothesis, the statistic $F_{\text{ld}} \sim F_{r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}, r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}}$. However, since the weakly sparse set $\boldsymbol{S}$ is unknown, the value of $r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}$ and $r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}$ cannot be determined in practice. From assumption (3.5), as $s^* = o(n^\tau/\log(p))$, then $F_{r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}, r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}} = F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^\perp}} + o_{\mathbb{P}}(1)$. Thus, the statistic F_{EW} can also be compared to the reference distribution $F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^\perp}}$.

Despite being asymptotically equivalent to the Wald F-test, F_{EW} has an additional bias term of $o_{\mathbb{P}}(n^{\tau-1})$, which impacts the power of the testing procedure. This leads us to consider the following hypotheses testing problem; for any $\tau \in [1/2, 1]$, we consider the contiguous hypotheses

$$H_0 : \lambda_{\max}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^\intercal\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}) = 0, \quad H_1 : \lambda_{\max, r_{\mathbf{Z}\ominus\mathbf{W}}}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^\intercal\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}) = hn^{\tau-1}. \tag{3.2.3}$$

Example 2. Consider the setting of Example 1 with $\boldsymbol{\nu}$ corresponding to a single random effect and $\boldsymbol{\nu} \sim \mathcal{N}_q(\mathbf{0}_q, \sigma_\nu^2\mathbf{I}_q)$. Then, with $\tau = 1/2$, the above hypotheses becomes

$$H_0 : \sigma_\nu^2 = 0, \qquad\qquad H_1 : m\sigma_\nu^2 = hn^{-1/2},$$

which is a standard hypotheses testing problem, such as in the balanced one-way random effects model. In this model, in the low-dimensional setting, the rate of $\sqrt{n}$ is optimal.

Theorem 3.2. *Consider the model given by equation (3.1.1) and the hypotheses testing problem from equation (3.2.3). Assume further (3.1), (3.2*), (3.3*), (3.4), and (3.5) for any* $\tau \in [1/2, 1]$. *Fix a value of* $\delta > 0$. *Under the alternative hypothesis with* $h > 0$ *sufficiently large (not depending on* n*) and* $\alpha \geq 4\left(\sigma_\varepsilon^2 + \lambda_{max}\left(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^\intercal\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\right)\right)$, *the sum of type I and type II errors for the test statistic* $\varphi_{F,\delta}$ *is less than one.*

Remark. The above theorem implies that F_{EW} can distinguish at the classical parametric $\sqrt{n}$ rate if the model is in the ultra-sparse regime, $s^* = o(\sqrt{n}/\log(p))$. This sparsity rate is common in high-dimensional inference problems for low-dimensional parameters at the parametric rate; in particular, for high-dimensional linear models, a version of this rate is required (cf. Cai and Guo (2017) and Javanmard and Montanari (2018)). When the value of $\tau \in (1/2, 1]$, we are limited by the ability to remove the bias from the mean vector; in the setting where $\tau = 1/2$, we are limited by the noise level. This seems to suggest a trade-off between the sparsity and the achievable rate of separation.

This comparison with the linear models literature that the inferential procedure requires an additional factor of $\sqrt{n}$ for sparsity assumption appears to be consistent with the recent results by Li et al. (2019). In particular, their proposed estimator for the variance components requires a consistent estimator of $\boldsymbol{\beta}^*$. They show in Theorem 3.1 that the minimax rate for estimating $\boldsymbol{\beta}^*$ is $s^* \log(p/s^2)/\operatorname{tr}(\boldsymbol{\Sigma}_a^{-1})$, where $\boldsymbol{\Sigma}_a \in \mathbb{R}^{n\times n}$ is a proxy for the true covariance matrix of $\mathbf{y}$. Thus, this suggests that $\operatorname{tr}(\boldsymbol{\Sigma}_a^{-1}) \asymp n$, and they require $s^* \log(p)/n \to 0$ to consistently estimate the variance components.

Remark. Compared to the recent work of Li et al. (2019), who only suggest an asymptotic distribution for their variance components estimators, Theorem 3.1 also demonstrates that F_{EW} enjoys certain optimality properties. In addition to providing a distribution under the null hypothesis, Theorem 3.1 also demonstrates under a sparsity assumption, F_{EW} is asymptotically equivalent to the classical Wald F-test, which is known to enjoy certain optimality properties, such as uniformly most powerful unbiased and uniformly most powerful invariant unbiased in certain ANOVA models (cf. Mathew and Sinha (1988)). In addition, Lu and Zhang (2010) showed that the Wald F-test and likelihood ratio tests are equivalent for balanced one-way ANOVA models while Qeadan and Christensen (2020) showed that the Wald F-test renders the likelihood ratio test inadmissible in generalized split plot designs. Moreover, unlike Li et al. (2019), who assume a compatibility condition, our procedure imposes no such requirement on the design matrix $\mathbf{X}$.

We now turn our attention to the setting of sub-Gaussian errors. When $\tau > 1/2$, z_{EW} no longer has an asymptotic Gaussian distribution at the $\sqrt{n}$ rate since the variance dominates the signal. Therefore, in this setting, we only consider hypotheses testing problems as given in equation (3.2.3) with $\tau = 1/2$.

Theorem 3.3. *Consider the model given by equation (3.1.1) and the hypotheses testing problem from equation (3.2.3). Assume further (3.1), (3.2), (3.3), (3.5) for $\tau \le 1/2$, and (3.7). Under the null hypothesis, if $\alpha \ge 4K_\varepsilon$, then*

$$\sqrt{n} z_{EW} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2_{\varsigma,z}\right).$$

Remark. Compared to Theorem 3.1, Theorem 3.3 trades the Gaussian assumption for a sub-Gaussian assumption under a slightly stronger sparsity assumption in order to obtain an asymptotic distribution. From Theorem 3.2, F_{EW} exhibits a continuous tradeoff between sparsity and power, which does not hold for z_{EW}. This is a consequence of using a central limit theorem for z_{EW}, which requires scaling by $\sqrt{n}$. This implies that the bias should be $o(\sqrt{n})$ and the signal from the alternative should be $\Omega(n^{-1/2})$.

Finally, we end this section by considering the setting where the number of random effects remains bounded.

Theorem 3.4. *Consider the model given in equation (3.1.1) and the hypotheses testing problem from equation (3.2.1). Assume (3.1), (3.2*), (3.3*), (3.4), and (3.6). If $\alpha > 4K_{Z\nu+\varepsilon}$ and $\tilde{\alpha} > 16\max(\operatorname{diag}(\mathbf{\Sigma}_X), \sigma_\varepsilon^2)$, then*

$$\tilde{F}_{EW} = F_{ld} + o_{\mathbb{P}}(n^{\tau-1}).$$

3.3 Confidence Intervals for a Single Random Effect

3.3.1 Model and Motivation

In the previous section, we considered the problem of testing a collection of random effects. However, it is often of interest to construct confidence intervals for the variance of a particular random effect. Suppose that $\Psi = \sigma_\nu^2 \mathbf{I}_v$. In the low-dimensional setting, there have been many procedures suggested to construct confidence intervals, from likelihood based approaches to F-test inversions (for example, see Jiang (2007) for a non-exhaustive list). In this section, we deal with a confidence interval for a single variance component, which can easily be extended using a Bonferroni correction or similar procedures for simultaneous confidence intervals. Alternatively, we may also invert the F-statistic from Section 2 to obtain confidence intervals for parameters of the form $\sigma_\nu^2/\sigma_\varepsilon^2$, with such ratios being first studied by Hartley and Rao (1967).

Our high-dimensional approach is inspired by F-test inversion. However, instead of using the ratio, we again use the difference. Define

$$\mathbf{Q} \triangleq \begin{pmatrix} \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} - r_{\mathbf{Z}\ominus\mathbf{W}} r^{-1}_{(\mathbf{Z},\mathbf{W})^\perp} \mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})} & \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z} \\ \mathbf{Z}^\intercal \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} & \mathbf{Z}^\intercal \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z} \end{pmatrix}, \qquad \boldsymbol{\xi} \triangleq \begin{pmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\nu} \end{pmatrix}.$$

Then, expanding the statistic z_{EW} from Section 3.2, we have that

$$\begin{aligned}
z_{\text{EW}} &= \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 - r_{\mathbf{Z}\ominus\mathbf{W}} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 \\
&= \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 - r_{\mathbf{Z}\ominus\mathbf{W}} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(n^\tau) \\
&= \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}) - r_{\mathbf{Z}\ominus\mathbf{W}}^{1/2} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1/2} \mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(n^\tau) \\
&= \boldsymbol{\xi}^\top\mathbf{Q}\boldsymbol{\xi} + o_{\mathbb{P}}(n^\tau),
\end{aligned}$$

where the second equality follows from Lemma A3.1 in the supplement. A direct calculation shows that $\mathbb{E}\boldsymbol{\xi}^\top\mathbf{Q}\boldsymbol{\xi} = \sigma_\nu^2 \operatorname{tr}(\mathbf{Z}^\top\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})$. Then, with proper centering and scaling, we may apply a central limit theorem for quadratic forms under a mild condition on the matrix $\mathbf{Q}$.

3.3.2 Estimator

To estimate σ_ν^2, we consider $\hat{\sigma}_\nu^2$ defined by

$$\hat{\sigma}_\nu^2 \triangleq [\operatorname{tr}(\mathbf{Z}^\top\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\Big(\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 - r_{\mathbf{Z}\ominus\mathbf{W}} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2\Big).$$

By a direct calculation, it can be shown that

$$\begin{aligned}
\sigma_\varsigma^2 \triangleq \text{Var}(\boldsymbol{\xi}^\top\mathbf{Q}\boldsymbol{\xi}) = \kappa_\varepsilon \sum_{i=1}^n \mathbf{Q}_{i,i}^2 + \kappa_\nu \sum_{i=n+1}^{n+q} \mathbf{Q}_{i,i}^2 \\
+ 2\sum_{i\neq j} \mathbf{Q}_{i,j}^2 (\sigma_\varepsilon^2 \mathbb{1}_{1\le i\le n} + \sigma_\nu^2 \mathbb{1}_{n+1\le i\le n+q})(\sigma_\varepsilon^2 \mathbb{1}_{1\le j\le n} + \sigma_\nu^2 \mathbb{1}_{n+1\le j\le n+q}).
\end{aligned}$$

From the above, we see that the asymptotic distribution of $\hat{\sigma}_\nu^2$ depends on the second and fourth moments of ν and ε. To estimate the second moment of ε, we consider the estimator

$$\hat{\sigma}_\varepsilon^2 \triangleq r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2.$$

The problem of estimation of fourth moments requires some technical assumptions on the design, even in the low-dimensional setting. For simplicity, we only consider the setting of Gaussian mixed models and the balanced one-way ANOVA design, but we note that the arguments may be extended under suitable regularity on the design matrices $\mathbf{Z}$ and $\mathbf{W}$. In the setting Gaussian mixed models, the fourth moment is entirely determined by the second moment. For the setting of the balanced one-way ANOVA design with m observations per subject, we consider the estimator

$$\begin{aligned}
&\hat{\omega}_\varepsilon \triangleq q^{-1} m^2 \|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{EW}})\|_4^4 - 3(m-1)\hat{\sigma}_\varepsilon^4, \\
&\hat{\omega}_\nu \triangleq (mq)^{-1} \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{EW}})\|_4^4 - 6m^{-1}\hat{\sigma}_\varepsilon^2\hat{\sigma}_\nu^2 - m^{-3}\hat{\omega}_\varepsilon - 3m^{-3}(m-1)\hat{\sigma}_\varepsilon^4, \\
&\hat{\kappa}_\varepsilon \triangleq \hat{\omega}_\varepsilon - \hat{\sigma}_\varepsilon^4, \qquad\qquad \hat{\kappa}_\nu \triangleq \hat{\omega}_\nu - \hat{\sigma}_\nu^4.
\end{aligned}$$

In both settings, we obtain a plug-in estimator $\hat{\sigma}_\varsigma^2$ of σ_ς^2. By setting $\hat{\kappa}_\nu = 0$ and $\hat{\sigma}_\nu^2 = 0$, we obtain an estimator $\hat{\sigma}_{\varsigma,z}^2$ of $\sigma_{\varsigma,z}^2$ for Section 3.2.

Remark. The statistic $\hat{\sigma}_\nu^2$ is related to the classical analysis of variance method for estimating random effects. Consider the setting of a balanced one-way ANOVA model from Example 1 with $\boldsymbol{\mu} = \mathbf{0}_n$. Let

$$\begin{aligned}
MS_{Treatments} &= q^{-1}\|\mathbf{P}_{\mathbf{Z}}\mathbf{y}\|_2^2 = q^{-1}\|\mathbf{P}_{\mathbf{Z}}(\mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon})\|_2^2, \\
MS_{Error} &= (n-q)^{-1}\|\mathbf{P}_{\mathbf{Z}}^{\perp}\mathbf{y}\|_2^2 = (n-q)^{-1}\|\mathbf{P}_{\mathbf{Z}}^{\perp}\boldsymbol{\varepsilon}\|_2^2.
\end{aligned}$$

Then, the analysis of variance estimate is given by

$$\hat{\sigma}_{\nu,AOV}^2 \triangleq m^{-1}(MS_{Treatments} - MS_{Error}).$$

Now, note that $r_{\mathbf{Z}\ominus\mathbf{W}} = q$, $r_{(\mathbf{Z},\mathbf{W})^\perp} = (m-1)q$, $n = mq$, and $\mathrm{tr}(\mathbf{Z}_{\mathbf{Z}}^\top\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}^\top\mathbf{Z}) = mq$. From the calculations in Section 3.3.1, we have that

$$\begin{aligned}
\hat{\sigma}_\nu^2 &= (mq)^{-1}\left(\|\mathbf{P}_{\mathbf{Z}}(\mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon})\|_2^2 - (m-1)^{-1}\|\mathbf{P}_{\mathbf{Z}}^{\perp}\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(q^\tau)\right) \\
&= \hat{\sigma}_{\nu,AOV}^2 + o_{\mathbb{P}}(1).
\end{aligned}$$

Thus, the two statistics are asymptotically equivalent in the balanced one-way ANOVA setting.

3.3.3 Assumptions

In addition to the assumptions from Section 3.2, we need additional assumptions on the matrix $\mathbf{Q}$ and on the distribution of the random effects ν.

(3.8) The matrix $\mathbf{Q}$ satisfies

$$\frac{\lambda_{\max}(\mathbf{Q}^2)}{\mathrm{tr}(\mathbf{Q}^2)} \to 0.$$

(3.9) The vector $\boldsymbol{\nu} = \boldsymbol{\nu}_n$ satisfies

$$\inf_n \left\{ \left(\min_{i=1,\dots,q_n} \mathrm{Var}(\boldsymbol{\nu}_{n,i}) \right) \wedge \left(\min_{i=1,\dots,q_n} \mathrm{Var}(\boldsymbol{\nu}_{n,i}^2) \right) \right\} > 0,$$
$$\lim_{x\to\infty} \sup_n \left\{ \left(\max_{i=1,\dots,q_n} \mathbb{E}\left(\boldsymbol{\nu}_{n,i}^2 : |\boldsymbol{\nu}_{n,i}| > x \right) \right) \vee \left(\max_{i=1,\dots,q_n} \mathbb{E}\left(\boldsymbol{\nu}_{n,i}^4 : |\boldsymbol{\nu}_{n,i}| > x \right) \right) \right\} = 0.$$

Remark. Assumptions (3.8) and (3.9), along with (3.7), are used for a central limit theorem for quadratic forms. For a thorough discussion on these assumptions, we refer the interested reader to Section 5 of Jiang (1996). As a consequence of using a central limit theorem, we require that the number of random effects increases to infinity. Thus, we only consider the sparsity assumption (3.5) in this section.

Example 3 (Balanced one-way ANOVA (ctd.))**.** Continuing with Example 1, we note that $\mathbf{Z}\mathbf{Z}^\mathsf{T} = m\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}$. Also, recall that $r_{\mathbf{Z}\ominus\mathbf{W}} = q$ and $r_{(\mathbf{Z},\mathbf{W})^\perp} = (m-1)q$. Then,

$$\mathbf{Q}^2 = \begin{pmatrix} (m+1)\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} + (m-1)^{-2}\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})} & (m+1)\mathbf{Z} \\ (m+1)\mathbf{Z}^\mathsf{T} & (m^2+m)\mathbf{I}_q \end{pmatrix}.$$

A direct calculation shows that $\lambda_{\max}(\mathbf{Q}^2) = (m+1)^2$ and $\mathrm{tr}(\mathbf{Q}^2) = (m+1)q + (m-1)^{-1}q + (m^2+m)q$, which satisfies assumption (3.8).

3.3.4 Main Results

We start by stating the asymptotic distribution of $\hat{\sigma}_\nu^2$.

Theorem 3.5. *Consider the model in equation (3.1.1). Assume (3.1), (3.2), (3.3) with* $\boldsymbol{\Psi} = \sigma_\nu^2 I_q$*, (3.4), (3.5) with* $\tau = 1/2$*, (3.7), (3.8), and (3.9). If* $\alpha > 4(K_\nu\,\lambda_{max}(\mathbf{Z}\mathbf{Z}^\mathsf{T}) + K_\varepsilon)$*, then*

$$\sigma_\varsigma^{-1}[\mathrm{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})](\hat{\sigma}_\nu^2 - \sigma_\nu^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Next, we consider the following lemma, which shows that $\hat{\kappa}_\varepsilon$ and $\hat{\kappa}_\nu$ are consistent estimators of κ_ε and κ_ν.

Proposition 3.6. *Consider the balanced one-way ANOVA from Example 1. Under the assumptions of Theorem 3.5,*

$$\hat{\kappa}_\varepsilon \xrightarrow{\mathbb{P}} \kappa_\varepsilon, \qquad \hat{\kappa}_\nu \xrightarrow{\mathbb{P}} \kappa_\nu.$$

Thus, the preceding two results allow us to construct confidence intervals in the Gaussian mixed model and the balanced one-way ANOVA setting. Let $\hat{\sigma}_\varsigma^2$ be a consistent estimator for σ_ς^2. Then, an asymptotic $(1-\delta)$ confidence interval for σ_ν^2 may be given by

$$(\hat{\sigma}_\nu^2 - z_\delta \hat{\sigma}_\varsigma [\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} \mathbf{Z})]^{-1}, \hat{\sigma}_\nu^2 + z_\delta \hat{\sigma}_\varsigma [\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} \mathbf{Z})]^{-1})$$

where z_δ is the δ upper quantile of a standard Gaussian distribution. Since the above interval may be negative, we may truncate negative values to zero.

3.4 Empirical Bayes in ANOVA Type Models

The motivating example of this problem framework is in terms of the Rasch model, originally proposed by Rasch (1960). The model that we consider is different than the classical Rasch model in that we have Gaussian responses as opposed to binary responses. Our interest in this section is not in testing whether the variance of the random effect is different from zero, but, assuming that it is different from zero, in estimating the individual components of the random effect. We use the term empirical Bayes, or compound decision, in the sense of Efron (2019) and the references therein (specifically Greenshtein and Ritov (2019)).

As an example of this model, the data that we consider in Section 3.6 is from the Trends in Mathematics and Sciences Study (TIMSS), an international study conducted every four years to measure fourth and eighth grade student achievement in mathematics and science. We only consider data from the year 2015. Polities randomly sample a collection of nationally representative schools to take standardized examinations in both mathematics and science, with questions being either multiple choice or constructed response. Then, each student within schools takes only a subset of the questions on the exams but all questions are answered by some students in each school. In addition to recording student responses, the data also contains background covariates for schools. Martin et al. (2016) provides a more detailed description of the methods and procedures employed by TIMSS and more general information about TIMSS is available in Mullis et al. (2016b).

For our analysis, we only consider multiple choice questions and analyze on the level of school rather than students. To construct a response variable for school, we compute the proportion of questions answered correctly by students in that school. Note that, unlike the classical Rasch model, we assume a linear model and, for all schools, we have answers for all questions. Thus, by a central limit theorem, our response $\mathbf{y}$ is approximately Gaussian. The fixed effects design $\mathbf{X}$ include the background covariates for the school and the random effects design $\mathbf{Z}$ is an indicator for the polity, with $\boldsymbol{\nu}$ corresponding to the unobserved variability of the polities. In this example, since we have averaged over questions, we do not have any nuisance random effects. The problem that we consider in this section is ranking the polities based on mathematical ability and trying to estimate the average number of questions that any particular polity will answer correctly. That is, we would like to estimate $\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\nu}$ for all polities in our data set.

3.4.1 Model and Motivation

The general problem framework that we consider is for K-factor ANOVA models. However, we derive the results in the setting when $K = 2$. That is, we consider the model

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

We do not assume that the design is fully crossed in the random effects. The goal in the problem is to estimate a subset of the mean vector, $\boldsymbol{\eta} \triangleq \boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\nu}$, since we view the random effects $\mathbf{W}$ as nuisance. However, as the sample size increases, the number of observations per group stays bounded. In the context of the motivating data example, each school still only answers a finite number of questions as we increase the sample size. A standard approach in the low-dimensional setting would be to use an empirical Bayes estimator by placing a Gaussian prior on both $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ (for example, see Brown et al. (2018)), which transforms the problem into a standard high-dimensional linear mixed model. Therefore, we use a $\mathcal{N}_v\left(\mathbf{0}_v, \sigma_\nu^2 \mathbf{I}_v\right)$ and $\mathcal{N}_r\left(\mathbf{0}_r, \sigma_\gamma^2 \mathbf{I}_r\right)$ prior on $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ respectively.

Since we need to estimate both σ_ν^2 and σ_γ^2 for the prior, our estimator for σ_γ^2 is analogous to $\hat{\sigma}_\nu^2$ from Section 3.3. To this end, we need an additional matrix $\mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}$ such that

$$\mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}\mathbf{y} = \mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}\mathbf{X}\boldsymbol{\beta}^* + \mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}\mathbf{W}\boldsymbol{\gamma} + \mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}\boldsymbol{\varepsilon}.$$

3.4.2 Estimator

Since we are also interested in estimating $\mathbf{P}_{\mathbf{W} \ominus \mathbf{Z}}\mathbf{X}\boldsymbol{\beta}^*$, we define $\tilde{\boldsymbol{\beta}}_{\text{EW}}$ to be the exponentially weighted estimator using the covariates $\mathbf{X}$, as opposed to using the covariates $\mathbf{P}_{\mathbf{W}}^{\perp}\mathbf{X}$ for $\hat{\beta}_{\text{EW}}$. Then, analogous to Section 3.2.2 let $\tilde{\boldsymbol{\beta}}_{\mathbf{m}}$ denote the least-squares estimator of $\boldsymbol{\beta}^*$ using the model $\mathbf{m} \in \mathcal{M}_u$ with covariates $\mathbf{X}_{\mathbf{m}}$ and $K_{\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}}$ be the sub-Gaussian parameter for $\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$. For $\tilde{\alpha} > 4K_{\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}}$, defining the exponential weights as

$$\tilde{w} \triangleq \frac{\exp\left(-\frac{1}{\tilde{\alpha}}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}\|_2^2\right)}{\sum_{\mathbf{k} \in \mathcal{M}_u} \exp\left(-\frac{1}{\tilde{\alpha}}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\mathbf{k}}\|_2^2\right)},$$

we have

$$\tilde{\boldsymbol{\beta}}_{\text{EW}} \triangleq \sum_{\mathbf{m} \in \mathcal{M}_u} \tilde{w}_{\mathbf{m}} \tilde{\boldsymbol{\beta}}_{\mathbf{m}}.$$

For convenience, we write $\tilde{\boldsymbol{\mu}}_{\text{EW}} \triangleq \mathbf{X}\tilde{\boldsymbol{\beta}}_{\text{EW}}$. Now, the estimators for the variance are given by

$$\tilde{\sigma}_\nu^2 \triangleq [\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\Big(\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\|_2^2 - r_{\mathbf{Z}\ominus\mathbf{W}} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\|_2^2\Big),$$
$$\tilde{\sigma}_\gamma^2 \triangleq [\operatorname{tr}(\mathbf{W}^\mathsf{T}\mathbf{P}_{\mathbf{W}\ominus\mathbf{Z}}\mathbf{W})]^{-1}\Big(\|\mathbf{P}_{\mathbf{W}\ominus\mathbf{Z}}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\|_2^2 - r_{\mathbf{W}\ominus\mathbf{Z}} r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\|_2^2\Big),$$
$$\tilde{\sigma}_\varepsilon^2 \triangleq r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\|_2^2.$$

As we do not require an asymptotic distribution for $\tilde{\sigma}_\nu^2$ and $\tilde{\sigma}_\gamma^2$, under weaker assumptions that Theorem 3.5, we have that $\tilde{\sigma}_\nu^2$ and $\tilde{\sigma}_\gamma^2$ are consistent estimators of σ_ν^2 and σ_γ^2 respectively. This suggests the the following empirical Bayes estimator for η,

$$\tilde{\boldsymbol{\eta}}_{\text{EW}} \triangleq \tilde{\boldsymbol{\mu}}_{\text{EW}} + \tilde{\sigma}_\nu^2\mathbf{Z}\mathbf{Z}^\mathsf{T}\left(\tilde{\sigma}_\nu^2\mathbf{Z}\mathbf{Z}^\mathsf{T} + \tilde{\sigma}_\gamma^2\mathbf{W}\mathbf{W}^\mathsf{T} + \tilde{\sigma}_\varepsilon^2\mathbf{I}_n\right)^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}}).$$

To compare our estimator, we consider an oracle that has access to $\boldsymbol{\mu}$, σ_ν^2, σ_γ^2, and σ_ε^2. Then, this oracle uses the Bayes estimator for $\boldsymbol{\eta}$ (see Lemma 3.8), given by

$$\tilde{\boldsymbol{\eta}}_{\text{oracle}} \triangleq \boldsymbol{\mu} + \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\mathsf{T}\left(\sigma_\nu^2\mathbf{Z}\mathbf{Z}^\mathsf{T} + \sigma_\gamma^2\mathbf{W}\mathbf{W}^\mathsf{T} + \sigma_\varepsilon^2\mathbf{I}_n\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}).$$

3.4.3 Assumptions

As previously mentioned, we do not need to establish the asymptotic distribution of $\hat{\sigma}_\nu^2$, rather we only need the estimator to be consistent. Accordingly, we may weaken our assumptions to the following

(3.10) The designs $\mathbf{Z}$ and $\mathbf{W}$ satisfy $\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}) \asymp \operatorname{tr}(\mathbf{W}^\mathsf{T}\mathbf{P}_{\mathbf{W}\ominus\mathbf{Z}}\mathbf{W}) \asymp n$.

(3.11) The matrix $\mathbf{W}$ satisfies $\lambda_{\text{max}}(\mathbf{W}\mathbf{W}^\mathsf{T})$ being bounded.

Remark. Assumption (3.10) ensures that the component of the design for the random effects is sufficiently well balanced. This assumption in the presence of (3.4) implies the second half of (3.5). Note that $\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}) \leq \lambda_{\text{max}}(\mathbf{Z}\mathbf{Z}^\mathsf{T})\operatorname{tr}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}) = \lambda_{\text{max}}(\mathbf{Z}\mathbf{Z}^\mathsf{T})r_{\mathbf{Z}\ominus\mathbf{W}}$. Since $r_{\mathbf{Z}\ominus\mathbf{W}} \leq n$, $\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}) \asymp n$ and $\lambda_{\text{max}}(\mathbf{Z}\mathbf{Z}^\mathsf{T})$ being bounded imply that $r_{\mathbf{Z}\ominus\mathbf{W}} \asymp n$.

The other assumption (3.11) is analogous to (3.4).

3.4.4 Main Results

We start this section by noting that $\hat{\sigma}_\nu^2$, $\hat{\sigma}_\gamma^2$, and $\hat{\sigma}_\varepsilon^2$ are all consistent estimators under a weaker sparsity assumption than in Section 3.3. Since we no longer require an asymptotic distribution for the variance estimates, we only need the prediction rate to ensure consistency, which is the content of the ensuing proposition.

Proposition 3.7. *Consider the model given in equation (3.1.1). Assume (3.1), (3.2*), (3.3*) with* $\mathbf{\Psi} = \sigma_\nu^2 \mathbf{I}_v$*, (3.4), (3.5) with* $\tau = 1$*, (3.10), and (3.11). If* $\alpha > 4(\sigma_\nu^2\, \lambda_{max}(\mathbf{Z}\mathbf{Z}^\mathsf{T}) + \sigma_\gamma^2\, \lambda_{max}(\mathbf{W}\mathbf{W}^\mathsf{T}) + \sigma_\varepsilon^2)$*, then*

$$\tilde{\sigma}_\nu^2 \xrightarrow{\mathbb{P}} \sigma_\nu^2, \qquad \tilde{\sigma}_\gamma^2 \xrightarrow{\mathbb{P}} \sigma_\gamma^2, \qquad \tilde{\sigma}_\varepsilon^2 \xrightarrow{\mathbb{P}} \sigma_\varepsilon^2.$$

The following is a standard lemma regarding the empirical Bayes estimators in this problem setup, which we prove for the sake of completeness.

Lemma 3.8. *For a fixed vector* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and fixed values* $\sigma_\nu^2 > 0$, $\sigma_\gamma^2 > 0$*, and* $\sigma_\varepsilon^2 > 0$*, the Bayes estimator of* $\boldsymbol{\eta}$ *is given by*

$$\mathbb{E}\left(\boldsymbol{\eta}|\mathbf{y}\right) = \boldsymbol{\mu} + \sigma_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T} \left(\sigma_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T} + \sigma_\gamma^2 \mathbf{W}\mathbf{W}^\mathsf{T} + \sigma_\varepsilon^2 \mathbf{I}_n\right)^{-1} (\mathbf{y} - \boldsymbol{\mu}).$$

We conclude this section with the main result regarding $\tilde{\boldsymbol{\eta}}_{\text{EW}}$; the empirical Bayes estimator performs nearly as well as the oracle Bayes estimator $\tilde{\boldsymbol{\eta}}_{\text{oracle}}$ asymptotically.

Theorem 3.9. *Consider the model given in equation (3.1.1). Under the assumptions of Proposition 3.7,*

$$n^{-1}\left(\|\tilde{\boldsymbol{\eta}}_{EW} - \boldsymbol{\eta}\|^2 - \|\tilde{\boldsymbol{\eta}}_{oracle} - \boldsymbol{\eta}\|^2\right) = o_{\mathbb{P}}(1).$$

3.5 Simulations

3.5.1 Methods and Models

We consider the linear mixed model given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

with $n = 1000$, $p = 2000$, and $q = 200$. The parameters that we vary throughout the experiment are the sparsity s, the distribution of $\mathbf{X}$, $\boldsymbol{\nu}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\varepsilon}$, the value of σ_ν^2, and the number of nuisance random effects d. For each parameter setting, the results are averaged over 100 replications.

For the sparsity, we set $s^* \in \{3, 15\}$. Each row of $\mathbf{X}$ is independent and identically distributed $\mathcal{N}_p\left(\mathbf{0}_p, \mathbf{\Sigma}\right)$ with

$$\mathbf{\Sigma}_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ \rho & \text{if } i \neq j, \end{cases}$$

for $\rho \in \{0, 0.8\}$. Then, $\boldsymbol{\beta}^*$ is chosen such that the signal strength, $(\boldsymbol{\beta}^*)^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta}^*$, is four times the noise level with $\sigma_\varepsilon^2 = 1$. This is accomplished by first generating s uniform random variables in $[-1, 1]$ and then rescaling to the desired level.

For the random effects, we either generate them from a Gaussian distribution or a double exponential distribution, which we denote by "z" and "e" respectively. For the variances, we let $\sigma_\nu^2 \in \{0, 1\}$ while $\sigma_\gamma^2 = 1$.

Finally, for the component of the design corresponding to the random effects, we let $d \in \{0, 200\}$. When $d = 0$, the design is a balanced one-way ANOVA design with $m = 5$. When $d = 200$, we generate from a two-way crossed design and down sample to have n observations. We only consider the sub-Gaussian procedures when $d = 0$.

All of our simulations are conducted in R. For each of our three problems, we compare the exponential weighting estimator, denoted by "EW" with an oracle low-dimensional estimator as well as a low-dimensional version of our proposed high-dimensional statistic.

For exponential weighting, we follow Algorithm 1 from Law and Ritov (2021b). Regarding the tuning parameters, we perform four fold cross-validation over a grid of values for α and the sparsity.

For the oracle estimators, in the setting of the F-test, we directly apply the classical low-dimensional F-test that has access to the true sparse set S, as given in equation (2.3) of Jiang (2007). For the confidence intervals, we fit the linear mixed models with the true sparse set S using `lmer` and applying the `confint` function. Finally, in the setting of estimation, we directly compute the oracle Bayes estimator $\tilde{\boldsymbol{\eta}}_{\text{oracle}}$ described in Section 3.4. Collectively, these low-dimensional estimators are denoted by "LD".

In addition to comparing with the low-dimensional estimators, we also construct low-dimensional versions of our proposed high-dimensional statistics. To do so, we use the exact same statistic as in the high-dimensional setting but replace exponential weighting with least-squares using the sparse set S. We make this comparison since all of our proposed statistics rely on two layers of asymptotics:

1. In the prediction of the mean vector via exponential weighting.

2. In the convergence once the residuals are obtained.

To differentiate between these two, we introduce an intermediate statistic that relies on least-squares, which we think of as low-dimensional versions of our statistics. For example, letting $\hat{\beta}^*{}_S$ be the least-squares estimator of β^* using the covariates X_S, we also consider the statistic

$$F_{\text{LS}} \triangleq \frac{\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{S}})\|^2 / r_{\mathbf{Z}\ominus\mathbf{W}}}{\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{S}})\|^2 / r_{(\mathbf{Z},\mathbf{W})^\perp}} \sim F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^\perp}} + o_{\mathbb{P}}(1).$$

These estimators are denoted by “LS”.

Finally, we also include a version of our statistics using scaled lasso, which we denote by “SL”. Then, for “EW”, “SL”, and “LS”, we subscript them by either “G” or “SG” to distinguish between the Gaussian and sub-Gaussian methods.

To compare the procedures, we consider the following metrics

1. Type I/II Error: The percentage of time the procedure produces a type I or type II error in hypothesis testing.

2. Average Coverage: The percentage of time the correct hypothesis is selected for F-tests or the percentage of time the true value of σ_ν^2 is in the confidence interval.

3. Average Length: The average length of the confidence interval, taken as the upper endpoint minus the lower endpoint.

4. Average Loss: The average squared Euclidean distance between the estimated vector $\hat{\boldsymbol{\eta}}$ and the true vector $\boldsymbol{\eta}$ divided by n.

3.5.2 Results

The results are presented in Tables A.3.1 – A.3.3 from the Supplement. We notice that for hypothesis testing, all the procedures control type I and type II error well throughout the settings. For confidence intervals, when $d = 0$ and $s = 3$, we notice that all of the methods perform well in coverage. However, the length of our procedures appears to be shorter when $\sigma_\nu^2 = 0$ and longer when $\sigma_\nu^2 = 1$, whereas the low-dimensional procedure is more uniform across the parameter space. This is not surprising in view of our estimation procedure. From Section 3.3.2, the asymptotic variance of $\hat{\sigma}_\nu^2$ depends monotonically on the second and fourth moments of ν and ε, which is reflected in the lengths of the resulting intervals.

When $\sigma_\nu^2 = 0$, the empirical coverage of our confidence intervals are close to the nominal level, even when the distribution of the random effects and errors are double exponential. When $\sigma_\nu^2 = 1$, the empirical coverage drops to around 80% for the Gaussian procedure and 90% for the sub-Gaussian procedure when the distribution is double exponential, against a nominal coverage of 95%. We note that the double exponential distribution is not a sub-Gaussian distribution, which seems to suggest that the confidence intervals are somewhat robust to slight departures from the distributional assumptions.

Moreover, when increasing the sparsity from $s^* = 3$ to $s^* = 15$, the performance of our confidence intervals decreases slightly since it is harder to remove the contribution of the fixed effects. Finally, for empirical Bayes estimation, our methods are competitive with the oracle. However, we notice that exponential weighting outperforms scaled lasso when $s^* = 15$, particularly when $\rho = 0.8$. Since larger values of ρ implies that the columns of X are more correlated, this highlights a salient feature of exponential weighting.

3.6 Real Data Application

Following in the motivating example of Section 3.4, we consider the TIMSS dataset, which is freely available at https://timssandpirls.bc.edu/. To simplify our analysis, we only consider the mathematics questions. After filtering out for complete cases on background covariates, we are left with 146 questions, $q = 43$ unique polities, $p = 106$ covariates, and 6808 schools. Therefore, we had a total of $n = 6808$ responses after averaging over the students and questions within the schools. Here, there are no nuisance random effects so $d = 0$. Due to averaging over students within schools, we expect the distributions to be approximately Gaussian by a central limit theorem.

To demonstrate our methodology, we use both exponential weighting as well as scaled lasso as our estimation procedure. When applying exponential weighting, we jointly tune the value of u and α using four fold cross-validation. The high-dimensional F-test rejected the null hypothesis that $\sigma_\nu^2 = 0$ and a 95% confidence interval for σ_ν^2 is $(0.0021, 0.0056)$, which suggests that, even controlling for school background characteristics, the polity of the school impacts mathematical ability. For the last part, we define a polity's background characteristics X to be the arithmetic average of all the schools' background characteristics within that polity. Then, applying the empirical Bayes procedure, we rank the polities based on the predicted number of questions they would answer correctly. The top five polities in order from our analysis are South Korea, Singapore, Hong Kong, Chinese Taipei, and Japan. Up to some reordering, our results are mostly consistent with the report of Mullis et al. (2016a) based on individual student data, who had the same top five polities. The results using scaled lasso produced the same ranking as exponential weighting and similar conclusions regarding σ_ν^2.

CHAPTER 4

High-Dimensional Varying Coefficient Models with Functional Random Effects

4.1 Introduction

Consider the following varying coefficients model with functional random effects given by

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j}) \rangle_2 + \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}) \tag{4.1.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Here, $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of time invariant covariates, $\mathbf{z}_i(\cdot) : (0,1) \to \mathbb{R}^q$ is a vector function representing the time varying covariates, and $\boldsymbol{\beta}^*(\cdot) : (0,1) \to \mathbb{R}^p$ and $\boldsymbol{\gamma}^*(\cdot) : (0,1) \to \mathbb{R}^q$ are time varying coefficients. Moreover, $\xi_i(\cdot) : (0,1) \to \mathbb{R}$ is a continuous time mean zero stochastic process representing the individual random effect and $\varepsilon_i(\cdot) : (0,1) \to \mathbb{R}$ is an independent error. Finally, the values $t_{i,j} \in (0,1)$ are the sampling times.

The model in (4.1.1) is useful for longitudinal data, where for the ith individual, we record m_i observations over time. Traditionally, varying coefficients models for longitudinal data consider errors that are mean zero stochastic process with unknown covariance structure, such as Hoover et al. (1998), whereas we partition the error into continuous time individual random effects, $\xi_i(t_{i,j})$, and independent errors, $\varepsilon_i(t_{i,j})$. Such a partitioning is reasonable whenever the mean function, $\xi_i(\cdot)$, for each individual is smooth. The model in equation (4.1.1) admits many special cases both in the low-dimensional setting and the ultra high-dimensional setting; we list two such examples below:

1. Let $p = 1$, $q = 0$, and $x_i = 1$ for all $i = 1, \ldots, n$, yielding the model

   $$y_i(t_{i,j}) = \beta^*(t_{i,j}) + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}). \tag{4.1.2}$$

   Then, the problem of estimating $\beta^*(\cdot)$ is equivalent to the mean function estimation problem from Cai and Yuan (2011).

2. Let $p > n$, $q = 0$, and $m_i = 1$ for all $i = 1, \ldots, n$, yielding the model

$$y_i(t_i) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_i) \rangle_2 + \varepsilon_i(t_i). \tag{4.1.3}$$

Here, in the model, since we only have a single observation per individual, we slightly abuse notation and write $\varepsilon_i(t_i)$ to denote $\xi_i(t_i) + \varepsilon_i(t_i)$. If the function $\boldsymbol{\beta}^*(\cdot)$ is sparse, then this corresponds to the sparse varying coefficients model considered by Klopp and Pensky (2015).

The goal of the present paper is estimation and inference for the varying coefficients $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$. Much of the extant literature on high-dimensional varying coefficient models focus on estimation and variable selection, such as Wei et al. (2011), Lian (2012), Xue and Qu (2012), Klopp and Pensky (2015), and Lee et al. (2016). The problem of inference is less well understood in the high-dimensional setting. To the best of our knowledge, the only paper exploring this problem is Chen and He (2018), who only consider local hypothesis testing. However, all of the current works assume that the time varying covariates, $\mathbf{z}_i(\cdot)$, are independent of the functional random effects, $\xi_i(\cdot)$. In practice, this assumption is not necessarily satisfied. To motivate this, consider modeling the average height of five year old boys in different countries over time. For each country, our time varying covariates consist of variables, such as Human Development Index (HDI), health expenditure, and urbanization rate, that are likely to be correlated with the functional random effect of country, which encapsulates, among other things, environmental factors. This example is revisited in Section 4.6. Moreover, most of the current works in high-dimensions assume either independent errors (cf. Wei et al. (2011), Xue and Qu (2012), and Klopp and Pensky (2015)) or a single individual, ie. $n = 1$ but $m \to \infty$ (cf. Lee et al. (2016)). The work most similar to ours is Bai et al. (2019), who assume m_i observations for individual, $i = 1, \ldots, n$. However, they only consider random and independent sampling times with correlated Gaussian errors in a Bayesian paradigm.

Thus, our contribution to high-dimensional varying cofficient models is fourfold: (i) estimation of $\boldsymbol{\beta}^*(\cdot)$ in the presence of $\xi_i(\cdot)$, (ii) estimation of $\boldsymbol{\gamma}^*(\cdot)$ under dependence of $\mathbf{z}_i(\cdot)$ and $\xi_i(\cdot)$, (iii) estimation of both $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$ under both random and independent or fixed and common sampling times, and (iv) construction of confidence bands for single components of $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$ with and without dependence.

To these ends, we propose a framework for estimation and inference in both the low-dimensional and the high-dimensional setting. Our estimators utilize orthogonal series to leverage recent developments in the high-dimensional linear models literature. We revisit the two examples in equations (4.1.2) and (4.1.3) later when analyzing the convergence rate of our estimators.

4.1.1 Organization of the Chapter

We end this section with a description of the notation that is used in the remainder of the chapter. Since our estimator has two-stages, we consider each stage separately. In Section 4.2, we consider the special case when there are no time invariant covariates ($p = 0$) while in Section 4.3, we consider the setting where there are no time varying covariates ($q = 0$). Then, we combine these results together into the two-stage estimator in Section 4.4. In Section 4.5, we consider the problem of constructing confidence bands for a fixed varying coefficient. Finally, Section 4.6 provides the simulations and Section 4.7 presents an analysis of the height data mentioned above. For ease of presentation, we defer all of the proofs to Appendix 3. In addition, the appendix contains the assumptions for Section 4.3, the additional results of the simulations, and an analysis of yeast cell cycle data.

4.1.2 Notation and Definitions

Throughout, all of our variables have a dependence on n, but, when it does not cause confusion, we suppress this dependence. Since our interest is mainly asymptotic, we adopt a random design regression framework embedded in a triangular array. The vector of covariates $\mathbf{x}_i$ is drawn from a distribution that does not depend on time whilst the vector of covariates $\mathbf{z}_i(t_{i,j})$ is drawn from a distribution conditioned on the sampling times. Furthermore, the $\xi_i(\cdot)$s are independent realizations of smooth mean zero stochastic processes and $\varepsilon_i(\cdot)$s are random errors. Then, we write $\mathbb{E}$ to denote the expectation with respect to the joint probability measure of $(\xi_i(\cdot))_{i=1}^n$ and $(\varepsilon_i(\cdot))_{i=1}^n$.

Regarding the sampling times, there are two commonly used paradigms: (i) independent random sampling times and (ii) common fixed sampling times. In the first setting, we assume that the time points are all independently sampled from a distribution f on $(0, 1)$, which is bounded away from zero and infinity. In this setting, $\mathbb{E}_T$ denotes the expectation with respect to the sampling times $t_{i,j}$. On the other hand, for the common sampling times, we assume that $m_i = m$ and $t_{i,j} = j/m$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$, viewing the sampling times as deterministic. In either case, for a fixed value of $i = 1, \ldots, n$, we write $(t_{i,(j)})_{j=1}^{m_i}$ to denote the order statistics for $(t_{i,j})_{j=1}^{m_i}$.

As mentioned in the Introduction, we consider an orthogonal series estimator. For technical convenience, we use the trigonometric basis since the functions are uniformly bounded by $\sqrt{2}$. There are many definitions of the trigonometric basis, but we use the following definition as in Tsybakov (2008).

Definition 4.1.1. For $t \in (0, 1)$, the trigonometric basis functions, denoted by $(\varphi_k(\cdot))_{k=1}^{\infty}$, are given by

$$\varphi_k(t) \triangleq \begin{cases} 1, & k = 1 \\ \sqrt{2}\cos(\pi k t), & k = 2, 4, \dots \\ \sqrt{2}\sin(\pi(k-1)t), & k = 3, 5, \dots. \end{cases}$$

Occasionally, it is useful to view these as functions on the complex plane; we write $\mathfrak{i}$ to denote the imaginary unit. Later, to simplify notation, we assume that both the varying coefficients and random effects are in the same periodic Sobolev class with smoothness α, denoted by $\mathcal{W}^{\text{per}}(\alpha, R)$ (for example, see Definition 1.11 of Tsybakov (2008)). Then, the functions $\boldsymbol{\beta}^*(\cdot)$, $\boldsymbol{\gamma}^*(\cdot)$, and $\xi_i(\cdot)$ admit an expansion over the trigonometric basis. Let $\beth_k^* \in \mathbb{R}^p$ (Hebrew letter Bet), $\gimel_k^* \in \mathbb{R}^q$ (Hebrew letter Gimel), and $\mathfrak{o}_{i,k} \in \mathbb{R}$ (Hebrew letter Samek) for $k = 1, 2, \dots$ denote the Fourier coefficients of $\boldsymbol{\beta}^*(\cdot)$, $\boldsymbol{\gamma}^*(\cdot)$, and $\xi_i(\cdot)$ for $i = 1, \dots, n$. Then, we may write

$$\boldsymbol{\beta}^*(\cdot) = \sum_{k=1}^{\infty} \beth_k^* \varphi_k(\cdot), \qquad \boldsymbol{\gamma}^*(\cdot) = \sum_{k=1}^{\infty} \gimel_k^* \varphi_k(\cdot), \qquad \xi_i(\cdot) = \sum_{k=1}^{\infty} \mathfrak{o}_{i,k} \varphi_k(\cdot). \tag{4.1.4}$$

Let $\hat{\boldsymbol{\beta}}(\cdot)$, with expansion

$$\hat{\boldsymbol{\beta}}(t) = \sum_{k=1}^{\infty} \hat{\beth}_k \varphi_k(t),$$

denote an arbitrary estimator for $\boldsymbol{\beta}^*(\cdot)$. To evaluate $\hat{\boldsymbol{\beta}}$, we consider either integrated squared error (ISE) defined by

$$\text{ISE}(\hat{\boldsymbol{\beta}}) \triangleq \int_0^1 \left(\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*(t)\right)^2 dt$$

or mean integrated squared error (MISE), where $\text{MISE}(\hat{\boldsymbol{\beta}}) \triangleq \mathbb{E}_T\mathbb{E}(\text{ISE}(\hat{\boldsymbol{\beta}}))$. We use MISE for the low-dimensional estimators and bound ISE for high-dimensional estimators with high probability. By Parseval's Theorem, it follows that integrated squared error is equivalent to

$$\text{ISE}(\hat{\boldsymbol{\beta}}) = \sum_{k=1}^{\infty} \left\|\hat{\beth}_k - \beth_k^*\right\|_2^2. \tag{4.1.5}$$

It follows from Proposition 1.14 of Tsybakov (2008) that

$$\sum_{k=K_\beta+1}^{\infty} \|\beth_k^*\|_2^2 = \mathcal{O}(s_\beta^* K_\beta^{-2\alpha}). \tag{4.1.6}$$

Therefore, it suffices to estimate the Fourier coefficients up to a truncation level K_β to balance the bias-variance tradeoff. Similarly, for $\boldsymbol{\gamma}^*(\cdot)$, we find a truncation level K_γ. Thus, we may then define the low and high frequency components of the varying coefficient functions $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$ as

$$\underline{\boldsymbol{\beta}^*}(\cdot) \triangleq \sum_{k=1}^{K_\beta} \beth_k^* \varphi_k(\cdot), \qquad \overline{\boldsymbol{\beta}^*}(\cdot) \triangleq \sum_{k=K_\beta+1}^{\infty} \beth_k^* \varphi_k(\cdot),$$
$$\underline{\boldsymbol{\gamma}^*}(\cdot) \triangleq \sum_{k=1}^{K_\gamma} \daleth_k^* \varphi_k(\cdot), \qquad \overline{\boldsymbol{\gamma}^*}(\cdot) \triangleq \sum_{k=K_\gamma+1}^{\infty} \daleth_k^* \varphi_k(\cdot).$$

Like other works in high-dimensional statistics, sparsity plays a crucial role. For simplicity, we assume the setting of strong sparsity, whereby both $\boldsymbol{\beta}^*(\cdot)$ and $\boldsymbol{\gamma}^*(\cdot)$ have s_β^* and s_γ^* components that are nonzero. When considering the inferential problem, we need another notion of sparsity from van de Geer et al. (2014). Let $\boldsymbol{\Sigma}$ to denote the population covariance matrix of $\mathbf{x}_i$, $\boldsymbol{\Theta}$ the inverse of $\boldsymbol{\Sigma}$, and $s_\theta = \max_{j=1,\ldots,p} |\{k \neq j : \boldsymbol{\Theta}_{j,k} \neq 0\}|$. Thus, s_θ is the maximal sparsity when regressing a component of $\mathbf{x}_i$ against the remaining x_i's.

4.2 Estimation with No Time Invariant Covariates

In this section, we assume that $p = 0$. That is, the model we consider is

$$y_i(t_{i,j}) = \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}). \tag{4.2.1}$$

Since the processes $\mathbf{z}_i(\cdot)$ and $\xi_i(\cdot)$ may have arbitrary dependence, to remove the effect of $\xi_i(\cdot)$ from the model, we difference the observations that are sufficiently close. As the function $\xi_i(\cdot)$ is assumed to be smooth, the value of $\xi_i(t)$ is approximately constant in a small neighborhood of t. Let $h > 0$ be a bandwidth tuning parameter. For simplicity, we temporarily assume that m_i is even for each $i = 1, \ldots, n$. Then, we may define the set $\mathscr{A}_h$ as

$$\mathscr{A}_h \triangleq \left\{(i,j) : 1 \leq i \leq n, j \in \{1, 3, \ldots, m_i - 1\}, t_{i,(j+1)} - t_{i,(j)} < h\right\}.$$

Let $N \triangleq N_h = |\mathscr{A}_h|$. For $(i,j) \in \mathscr{A}_h$, define the differenced observations $v_{i,j}$ as

$$
\begin{aligned}
v_{i,j} &\triangleq y_i(t_{i,(j+1)}) - y_i(t_{i,(j)}) \\
&= \langle \mathbf{z}_i(t_{i,(j+1)}), \boldsymbol{\gamma}^*(t_{i,(j+1)}) \rangle_2 - \mathbf{z}_i(t_{i,(j)}), \boldsymbol{\gamma}^*(t_{i,(j)}) \rangle_2 + \xi_i(t_{i,(j+1)}) - \xi_i(t_{i,(j)}) \\
&\quad + \varepsilon_i(t_{i,(j+1)}) - \varepsilon_i(t_{i,(j)}) \\
&= \sum_{k=1}^{K_\gamma} \langle \underbrace{\varphi_k(t_{i,(j+1)})\mathbf{z}_i(t_{i,(j+1)}) - \varphi_k(t_{i,(j)})\mathbf{z}_i(t_{i,(j)})}_{\psi_{i,j,k}^{\mathsf{T}}}, \gimel_k^* \rangle_2 + \underbrace{\varepsilon_i(t_{i,(j+1)}) - \varepsilon_i(t_{i,(j)})}_{\eta_{i,j}} \\
&\quad + \underbrace{\sum_{k=K_\gamma+1}^{\infty} \langle \varphi_k(t_{i,(j+1)})\mathbf{z}_i(t_{i,(j+1)}) - \varphi_k(t_{i,(j)})\mathbf{z}_i(t_{i,(j)}), \gimel_k^* \rangle_2 + \xi_i(t_{i,(j+1)}) - \xi_i(t_{i,(j)})}_{\Delta_{i,j}^{(\gamma)}} \\
&= \boldsymbol{\psi}_{i,j}^{\mathsf{T}} \gimel^* + \eta_{i,j} + \Delta_{i,j}^{(\gamma)},
\end{aligned}
$$

where $\boldsymbol{\psi}_{i,j} = (\boldsymbol{\psi}_{i,j,1}^{\mathsf{T}}, \ldots, \boldsymbol{\psi}_{i,j,K_\gamma}^{\mathsf{T}})^{\mathsf{T}}$ and $\gimel^* = (\gimel_1^{*\mathsf{T}}, \ldots, \gimel_{K_\gamma}^{*}{}^{\mathsf{T}})^{\mathsf{T}}$. In matrix notation, we write this model as

$$
\mathbf{v} = \boldsymbol{\Psi}\gimel^* + \boldsymbol{\eta} + \boldsymbol{\Delta}^{(\gamma)}. \tag{4.2.2}
$$

This is a sparse high-dimensional partially linear model with uncorrelated errors, for which there are many proposals for estimating $\gimel^*$. Commonly, in high-dimensional nonparametric models, a version of group lasso (cf. Yuan and Lin (2006)) is used to select relevant functions after a basis expansion, such as the SpAM estimator of Ravikumar et al. (2009) or the block lasso estimator of Klopp and Pensky (2015). While such approaches lead to more interpretable estimators, we estimate $\gimel^*$ by the classical lasso of Tibshirani (1996) but note that our approach generalizes to using the group lasso. We use the classical lasso to motivate the inferential procedure in Section 4.5, which is based on a version of the de-biased lasso estimator. Therefore, we estimate $\gimel^*$ in the low-dimensional case by

$$
\hat{\gimel}^{\mathrm{LD}} \triangleq \left(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Psi}^{\mathsf{T}}\mathbf{v}
$$

and

$$
\hat{\gimel}^{\mathrm{HD}} \triangleq \operatorname*{arg\,min}_{\gimel \in \mathbb{R}^{qK_\gamma}} N^{-1} \left\| \mathbf{v} - \boldsymbol{\Psi}\gimel \right\|_2^2 + \lambda \left\| \gimel \right\|_1
$$

in the high-dimensional setting for a suitable tuning parameter $\lambda > 0$. By identifying the vectors $\hat{\gimel}^{\text{LD}}$ and $\hat{\gimel}^{\text{HD}}$ as K_γ vectors in $\mathbb{R}^q$, the estimators for $\boldsymbol{\gamma}^*(\cdot)$ are given by

$$\hat{\boldsymbol{\gamma}}^{\text{LD}}(\cdot) \triangleq \sum_{k=1}^{\infty} K_\gamma \hat{\gimel}_k^{\text{LD}} \varphi_k(\cdot), \qquad \hat{\boldsymbol{\gamma}}^{\text{HD}}(\cdot) \triangleq \sum_{k=1}^{\infty} K_\gamma \hat{\gimel}_k^{\text{HD}} \varphi_k(\cdot).$$

Remark. In the above formulation, we pair the observations in $\mathscr{A}_h$ to ensure that the resultant errors $\boldsymbol{\eta}$ are independent. However, this reduces the number of observations that we have to estimate $\gimel^*$. To circumvent this problem, we may alternatively consider the set

$$\mathscr{B}_h \triangleq \left\{(i,j) : 1 \le i \le n, 1 \le j \le m_i - 1, t_{i,(j+1)} - t_{i,(j)} < h\right\}.$$

Using the set $\mathscr{B}_h$, we may likewise form the model given in equation (4.2.2), where the resultant partially linear model has, by construction, correlated errors with known correlation structure. Hence we may find a matrix $\mathbf{B}$ such that

$$\mathbf{B}\mathbf{v} = \mathbf{B}\boldsymbol{\Psi}\gimel^* + \mathbf{B}\boldsymbol{\eta} + \mathbf{B}\boldsymbol{\Delta}^{(\gamma)},$$

where $\mathbf{B}\boldsymbol{\eta}$ is uncorrelated. For simplicity, we consider only the set $\mathscr{A}_h$, but in practice, we recommend adjusting using $\mathscr{B}_h$ when the sampling times are random and independent but $\mathscr{A}_h$ when the sampling times are fixed and common (see Table A.4.3 in Section A.4.6 of the Supplement).

4.2.1 Sample Size

In this subsection, we consider the expected number of observations after differencing under a few asymptotic regimes for n, m, and h. This leads us to the following proposition.

Proposition 4.1. *Suppose the sampling times* $t_{i,j} \overset{\text{i.i.d.}}{\sim} f$ *for a density* f *on* $(0,1)$ *bounded away from zero and infinity and* $m_i = m > 0$ *for all* $i = 1, \ldots, n$. *Let* $\tilde{N}_h = |\mathscr{B}_h|$.

1. *If* $m = \mathcal{O}(1)$ *and* $n \to \infty$, *then* $\mathbb{E}_T \tilde{N}_h \asymp nh$.

2. *If* $m \to \infty$ *and* $mh \ll 1$, *then* $\mathbb{E}_T \tilde{N}_h \asymp nm^2h$ *and* $\mathbb{E}_T(N_h + 1)^{-1} \asymp (nm^2h)^{-1}$.

3. *If* $m \to \infty$ *and* $mh \gg 1$, *then* $\mathbb{E}_T \tilde{N}_h \asymp nm$. *If, in addition,* $mh - \log(mn) \to \infty$, *then* $\mathbb{P}(\tilde{N}_h = mn) \to 1$.

Remark. It is easy to see that $|\mathscr{A}_h| \asymp |\mathscr{B}_h|$.

4.2.2 Assumptions

The following assumptions are used when $p = 0$.

(4.1) If $s_\gamma^* = q < N$, then the matrix $\mathbf{\Psi}$ satisfies $\mathrm{tr}[(\mathbf{\Psi}^\mathsf{T}\mathbf{\Psi})^{-1}] = \mathcal{O}(s_\gamma^* K_\gamma / N)$ and $\|(\mathbf{\Psi}^\mathsf{T}\mathbf{\Psi})^{-1}\|_2 = \mathcal{O}_\mathbb{P}(N^{-1})$.

(4.2) The columns of the matrix $\mathbf{\Psi}$ have squared norms that are uniformly $\mathcal{O}_\mathbb{P}(N)$.

(4.3) The design matrix $\mathbf{\Psi}$ satisfies the compatibility condition with compatibility constant $\phi_{\text{cc},\Psi} > 0$.

(4.4) The design matrix $\mathbf{\Psi}$ satisfies the adaptive restricted eigenvalue condition with constant $\phi_{\text{adap},\Psi} > 0$.

(4.5) The errors $\varepsilon_i(t_{i,j})$ are independent and identically distributed with mean zero and variance σ_ε^2. Moreover, the errors are independent of the sampling times.

(4.6) The errors $\varepsilon_i(t_{i,j}) \overset{\text{i.i.d.}}{\sim} \mathcal{SG}(\varsigma_\varepsilon^2)$. Moreover, the errors are independent of the sampling times.

(4.7) The functional random effects $(\xi_i(\cdot))_{i=1}^n$ are uniformly Lipschitz with constant L.

(4.8) Each coordinate of the coefficient $\boldsymbol{\gamma}^*(\cdot)$ satisfies $\boldsymbol{\gamma}_k^*(t) \in \mathcal{W}^{\text{per}}(\alpha, R)$ for some constant $\alpha \geq 2$ and $R > 0$ with s_γ^* coordinates nonzero. Moreover,

$$\mathbb{E}_T \left(\sum_{k=K_\gamma+1}^{\infty} \langle \mathbf{z}_i(t), \mathbf{\daleth}_k^* \rangle_2 \varphi_k(t) \right)^2 = \mathcal{O}(s_\gamma^* K_\gamma^{-2\alpha}).$$

Remark. Assumptions (4.1), (4.2), and (4.5) are standard scaling assumptions for the design in the low-dimensional setting.

Assumptions (4.3) and (4.4) are both compatibility conditions on the design matrix, with (4.4) implying (4.3). In Theorem 4.3 below, we use (4.3) to obtain slow rates on ISE while (4.4) yields a fast rate on ISE. For a more detailed discussion on assumptions (4.3) and (4.4), including definitions, we refer the reader to Section 6.2 of Bühlmann and van de Geer (2011) and Section 4 of Bickel et al. (2009) respectively.

Next, assumption (4.6) is standard in the high-dimensional linear models literature and (4.7) is reasonable whenever the underlying mean function for each individual is smooth. Further, (4.7) is implied whenever $(\xi_i(\cdot))_{i=1}^n \subseteq \mathcal{W}^{\text{per}}(\alpha, R)$.

The first half of assumption (4.8) is standard in the literature on nonparametric regression while the second half ensures that the varying coefficients can be well approximated by a few basis functions. In Example 4 below, we consider an instance where the second half is satisfied.

Example 4 (Example for Assumption (4.8): Time random covariates)**.** Suppose that the distribution of $\mathbf{z}_i(t_{i,j})$ does not depend on $t_{i,j}$. That is, assume that $\mathscr{L}\{\mathbf{z}_i(t_{i,j})|t_{i,j}\} = \mathscr{L}\{\mathbf{z}_i(t_{i,k})|t_{i,k}\}$ for every $j, k = 1, \ldots, m_i$ with $j \neq k$. Then,

$$\begin{aligned}
\mathbb{E}_T \left(\sum_{k=K_\gamma+1}^{\infty} \langle \mathbf{z}_i(t), \daleth_k^* \rangle_2 \varphi_k(t) \right)^2 &= \sum_{k=K_\gamma+1}^{\infty} \mathbb{E}_T \langle \mathbf{z}_i(t), \daleth_k^* \rangle_2^2 \\
&= \sum_{k=K_\gamma+1}^{\infty} \daleth_k^{*\top} \mathbb{E}_T \left([\mathbf{z}_i(t)][\mathbf{z}_i(t)]^\top \right) \daleth_k^* \\
&\leq \left\| \mathbb{E}_T \left([\mathbf{z}_i(t)][\mathbf{z}_i(t)]^\top \right) \right\|_2^2 \sum_{k=K_\gamma+1}^{\infty} \|\daleth_k^*\|_2^2 \\
&= \mathcal{O}\left(s_\gamma^* K_\gamma^{-2\alpha} \right).
\end{aligned}$$

4.2.3 Main Results

We start by stating a result for the low-dimensional setting.

Proposition 4.2. *Consider the model given in equation* (4.2.2)*. Assume (4.1), (4.5), (4.7), and (4.8). Then*

$$\mathbb{E}\mathbb{E}_T \left\| \hat{\daleth}^{LD} - \daleth^* \right\|_2^2 = \mathcal{O}\left(s_\gamma^* K_\gamma \mathbb{E}_T N^{-1} + s_\gamma^* K_\gamma^{-2\alpha} + L^2 h^2 \right).$$

Remark. As noted in Section 4.1.2, the MISE of $\hat{\boldsymbol{\gamma}}^{\mathrm{LD}}(\cdot)$ can be bounded by

$$\mathrm{MISE}(\hat{\boldsymbol{\gamma}}^{\mathrm{LD}}) = \mathcal{O}\left(s_\gamma^* K_\gamma \mathbb{E}_T N^{-1} + s_\gamma^* K_\gamma^{-2\alpha} + L^2 h^2 \right).$$

Choosing $K_\gamma \asymp (\mathbb{E}_T N^{-1})^{-1/(2\alpha+1)}$ yields

$$\mathrm{MISE}(\hat{\boldsymbol{\gamma}}^{\mathrm{LD}}) = \mathcal{O}\left(s_\gamma^* (\mathbb{E}_T N^{-1})^{2\alpha/(2\alpha+1)} + L^2 h^2 \right).$$

The choice of h is less straightforward as it depends on the asymptotic growth of m relative to n, which can be seen from Proposition 4.1.

Now, turning our attention to the high-dimensional setting, we have the following result.

Theorem 4.3. *Consider the model given in equation* (4.2.2)*. Assume (4.2), (4.6), (4.7), and (4.8). For $t > 0$, let*

$$\lambda_0 \triangleq 2\varsigma_\varepsilon \sqrt{N^{-1} \max_{j=1,\ldots,p} \|\boldsymbol{\Psi}_j\|_2^2} \sqrt{\frac{t^2 + 2\log(qK_\gamma)}{N}}. \tag{4.2.3}$$

Suppose $\lambda \geq 2\lambda_0$.

1. *If, in addition, (4.3) holds, then with probability at least* $1 - 2\exp(-t^2/2)$,

$$2N^{-1}\left\|\mathbf{\Psi}\hat{\daleth}^{HD} - \mathbf{\Psi}\daleth^{*} - \mathbf{\Delta}^{(\gamma)}\right\|_2^2 + \lambda\left\|\hat{\daleth}^{HD} - \daleth^{*}\right\|_1 \leq 6N^{-1}\left\|\mathbf{\Delta}^{(\gamma)}\right\|_2^2 + 24\phi_{cc,\Psi}^{-2}\lambda^2 s_\gamma^* K_\gamma.$$

2. *If, in addition to the above, (4.4) holds, then with probability at least* $1 - 2\exp(-t^2/2)$,

$$\left\|\hat{\daleth}^{HD} - \daleth^{*}\right\|_2^2 = \mathcal{O}\left(\lambda^2 s_\gamma^* K_\gamma \left(\frac{s_\gamma^* K_\gamma^{-2\alpha} + L^2h^2}{\lambda^2 s_\gamma^* K_\gamma} + \phi_{adap,\Psi}^{-2}\right)^2\right).$$

Example 5. As a special case, consider the setting where $n = 1$ and $m = m_1 \to \infty$. Note that this problem is a generalization of the model in equation (4.1.3) from the Introduction. Under the additional assumption that $\xi_i(\cdot) \equiv 0$, obtaining m observations from a single individual with time varying covariates is equivalent to obtaining a single observation from m individuals. Set $\lambda = 2\lambda_0$, $K_\gamma \asymp (m/\log(q))^{1/(2\alpha+1)}$ and $h \asymp s_\gamma^*(\log(q)/m)^{2\alpha/(2\alpha+1)}$. Since $q > K_\gamma$, it follows that $\log(q) \leq \log(qK_\gamma) \leq 2\log(q)$. Moreover, the choice of h implies that $mh \gg 1$; thus, $N = m$ with high probability for m sufficiently large by Proposition 4.1. Then, with probability at least $1 - 2\exp(-t^2/2)$, it follows from Theorem 4.3 that

$$\begin{aligned}\left\|\hat{\daleth}^{\text{HD}} - \daleth^{*}\right\|_2^2 &= \mathcal{O}\left(\frac{s_\gamma^* m}{K_\gamma^{4\alpha+1}\log(qK_\gamma)} + \frac{h^2 m}{s_\gamma^* K_\gamma \log(qK_\gamma)} + \frac{s_\gamma^* K_\gamma \log(qK_\gamma)}{m}\right)\\ &= \mathcal{O}\left(s_\gamma^*\left(\frac{\log(q)}{m}\right)^{2\alpha/(2\alpha+1)}\right).\end{aligned}$$

From equation (4.1.6), it follows that

$$\sum_{k=K_\gamma+1}^{\infty}\|\daleth_k^*\|_2^2 = \mathcal{O}\left(s_\gamma^* K_\gamma^{-2\alpha}\right).$$

Combining these facts yields the following bound on ISE with probability at least $1-2\exp(-t^2/2)$.

$$\text{ISE}(\hat{\boldsymbol{\gamma}}^{\text{HD}}) = \left\|\hat{\daleth}^{\text{HD}} - \daleth^{*}\right\|_2^2 + \sum_{k=K_\gamma+1}^{\infty}\|\daleth_k^*\|_2^2 = \mathcal{O}\left(s_\gamma^*\left(\frac{\log(q)}{m}\right)^{2\alpha/(2\alpha+1)}\right).$$

From Theorem 1 of Klopp and Pensky (2015), assuming that the functions have uniform smoothness α, then, up to the logarithmic factor, $\hat{\boldsymbol{\gamma}}^{\text{HD}}(\cdot)$ attains the minimax rate.

4.3 Estimation with No Time Varying Covariates

In this section, we assume that $q = 0$. That is, the model we consider is

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}). \tag{4.3.1}$$

Now, by directly substituting the expansions from equation (4.1.4), it follows that

$$y_i(t_{i,j}) = \sum_{k=1}^{\infty} \left(\langle \mathbf{x}_i, \beth_k^* \rangle_2 + \mathfrak{o}_{i,k} \right) \varphi_k(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

The above factoring suggests that, to estimate the kth Fourier coefficient $\beth_k^*$, we should look at the observations in the frequency domain as opposed to the time domain. That is, we should form new observations in the frequency domain as

$$\boldsymbol{\omega}_{i,k} \triangleq m_i^{-1} \sum_{j=1}^{m_i} y_i(t_{i,j}) \varphi_k(t_{i,j}).$$

By projecting the observations onto a fixed frequency given by φ_k, the above is an approximate linear model. Depending on whether the sampling times are viewed as fixed and common or random and independent, we partition the above model differently.

In the setting where the sampling times are random and independent for each individual, we define

$$\zeta_{i,k} \triangleq m_i^{-1} \sum_{j=1}^{m_i} \left(\langle \mathbf{x}_i, \beta^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}) \right) \varphi_k(t_{i,j}) - \langle \mathbf{x}_i, \beth_k^* \rangle_2,$$

yielding a linear model

$$\boldsymbol{\omega}_{i,k} = \langle \mathbf{x}_i, \beth_k^* \rangle_2 + \zeta_{i,k}. \tag{4.3.2}$$

In matrix notation, we write

$$\boldsymbol{\Omega}_k = \mathbf{X} \beth_k^* + \boldsymbol{\zeta}_k.$$

When the sampling times are fixed and common, we write

$$\boldsymbol{\omega}_{i,k} = \langle \mathbf{x}_i, \beth_k^* + \daleth_k \rangle_2 + \zeta_{i,k}, \tag{4.3.3}$$

where

$$\daleth_k \triangleq m^{-1} \sum_{\substack{l=1 \\ l \neq k}}^{\infty} \beth_l^* \sum_{j=1}^{m} \varphi_k(t_{i,j}) \varphi_l(t_{i,j}), \quad \text{(Hebrew letter Dalet)}$$

$$\zeta_{i,k} \triangleq \mathfrak{o}_{i,k} + m^{-1} \sum_{j=1}^{m} \varphi_k(t_{i,j}) \left(\varepsilon_i(t_{i,j}) + \sum_{\substack{l=1 \\ l \neq k}}^{\infty} \mathfrak{o}_{i,l} \varphi_l(t_{i,j}) \right).$$

In matrix notation, this is expressed as

$$\mathbf{\Omega}_k = \mathbf{X}(\beth_k^* + \daleth_k) + \boldsymbol{\zeta}_k.$$

The main difference between the perspectives in equations (4.3.2) and (4.3.3) is how we consider the inexact orthogonalization. When the sampling times are random, the projection of $\langle \mathbf{x}_i, \beta^*(\cdot) \rangle_2$ over the k'th frequency is unbiased for $\langle \mathbf{x}_i, \beth_k^* \rangle_2$ with respect to the probability measure on $t_{i,j}$. Conversely, since the common sampling times are deterministic, this inexact orthogonalization is viewed as bias; the vector $\daleth_k$ is the sum of the Fourier coefficients of the aliased frequencies.

Regardless of the sampling times, in the low-dimensional setting, least-squares provides a convenient estimator for $\beth_k^*$, while in the high-dimensional setting, one may directly apply the lasso. That is,

$$\hat{\beth}_k^{\text{LD}} \triangleq \begin{cases} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{\Omega}_k, & k = 1, \ldots, K_\beta, \\ \mathbf{0}_p, & k > K_\beta, \end{cases}$$

$$\hat{\beth}_k^{\text{HD}} \triangleq \begin{cases} \arg\min_{\beth^* \in \mathbb{R}^p} n^{-1} \left\| \mathbf{\Omega}_k - \mathbf{X}\beth^* \right\|_2^2 + \lambda_k \left\| \beth^* \right\|_1, & k = 1, \ldots, K_\beta, \\ \mathbf{0}_p, & k > K_\beta. \end{cases}$$

Like in Section 4.2, this provides the estimators for $\beta^*(\cdot)$ as

$$\hat{\boldsymbol{\beta}}^{\text{LD}}(\cdot) \triangleq \sum_{k=1}^{\infty} \hat{\beth}_k^{\text{LD}} \varphi_k(\cdot), \qquad \hat{\boldsymbol{\beta}}^{\text{HD}}(\cdot) \triangleq \sum_{k=1}^{\infty} \hat{\beth}_k^{\text{HD}} \varphi_k(\cdot).$$

4.3.1 Assumptions

We begin with the assumptions we use in this section.

(4.9) The sampling times satisfy $t_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ and are independent of the errors $\varepsilon_i(t_{i,j})$.

(4.10) The number of observations per individual is the same, $m_i = m$ for all $i = 1, \ldots, n$. Moreover, the sampling times satisfy $t_{i,j} = j/m$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

(4.11) The columns of $\mathbf{X}$ have squared norms that are uniformly $\mathcal{O}(n)$ and the entries of $\mathbf{X}$ satisfy $\sup_{i=1,\ldots,n} \|\mathbf{x}_i\|_\infty = g(n)$ for some function $g(n)$.

(4.12) The design matrix $\mathbf{X}$ satisfies the compatibility condition with compatibility constant $\phi_{\text{cc},X} > 0$.

(4.13) The design matrix $\mathbf{X}$ satisfies the adaptive restricted eigenvalue condition with constant $\phi_{\text{adap},X} > 0$.

(4.14) Each coordinate of the coefficient $\boldsymbol{\beta}^*(\cdot)$ satisfies $\boldsymbol{\beta}_k^*(\cdot) \in \mathcal{W}^{\text{per}}(\alpha, R)$ for some constant $\alpha \geq 2$ and $R > 0$ with s_β^* coordinates nonzero. Moreover, the product $\langle \mathbf{x}_i, \boldsymbol{\beta}^*(\cdot)\rangle_2 \in \mathcal{W}^{\text{per}}(\alpha, \mathcal{O}(g(n))$ uniformly and

$$\sup_{i=1,\ldots,n} \int_0^1 \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t)\rangle_2^2 dt = \mathcal{O}(g(n))$$

for some function $g(n)$.

(4.15) The individual random effects $(\xi_i(\cdot))_{i=1}^n \subseteq \mathcal{W}^{\text{per}}(\alpha, R)$ almost surely and are independent and identically distributed. Furthermore, for each $i = 1, \ldots, n$ and $t \in (0, 1)$, the function $\xi_i(\cdot)$ satisfies $\mathbb{E}\xi_i(t) = 0$ and

$$\mathbb{E} \int_0^1 \xi_i^2(t) dt < \infty.$$

(4.16) The individual random effects $(\xi_i(\cdot))_{i=1}^n \subseteq \mathcal{W}^{\text{per}}(\alpha, R)$ almost surely and are independent and identically distributed. Furthermore, for each $i = 1, \ldots, n$, the Fourier coefficients $\mathfrak{o}_{i,k} \sim \mathcal{SG}(\varsigma_{\mathfrak{o},k}^2)$ with $\sum_{k=K}^\infty \varsigma_{\mathfrak{o},k}^2 = \mathcal{O}(K^{-2\alpha})$.

There are two assumptions regarding the sampling times, (4.9) and (4.10). As first pointed out by Cai and Yuan (2011) in the context of mean function estimation, the rate of convergence is different between random and independent or fixed and common sampling times. The first assumption, (4.9), considers the independent sampling times, with an additional convenience that the sampling times are uniform since the trigonometric basis is orthogonal with respect to this measure on $(0, 1)$. This assumption may be relaxed to allow for $t_{i,j} \overset{\text{i.i.d.}}{\sim} f$ from some density f bounded from zero and infinity by taking an appropriate change of measure. The other assumption, (4.10), considers the common sampling time setting. Again, we make a simplifying assumption that the sampling times are on a uniformly spaced grid, though this assumption may similarly be relaxed. These two settings are analyzed separately below.

Assumptions (4.11) – (4.13) are analogous (4.2) – (4.4).

Next, assumption (4.14) assumes that the signal is uniformly bounded by some function $g(n)$. This assumption is the finite sample analogue of maintaining a bounded signal to noise ratio in a varying coefficients model. We conflate the function $g(n)$ in assumptions (4.11) and (4.14) since $g(n)$ may normally be taken to be a slowly varying function of n. For example, if the design is sub-Gaussian, then $g(n) = \log(n)$ is sufficient. If the design is bounded, then $g(n)$ may be a constant. Further, we note that this assumption implies that

$$\sup_{i=1,\ldots,n} \sup_{t\in(0,1)} \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t)\rangle_2^2 dt = \mathcal{O}(g(n)).$$

Assumption (4.15) is a slightly stronger version of (4.7). It automatically implies that $\mathbb{E}\mathfrak{o}_{i,k} = 0$ for all $k = 1, 2, \ldots$, and $\sum_{k=K_\beta}^{\infty} \sigma_{\mathfrak{o},k}^2 = \mathcal{O}(K_\beta^{-2\alpha})$ where $\sigma_{\mathfrak{o},k} = \mathrm{Var}(\mathfrak{o}_{i,k})$. Finally, (4.16) is a technical requirement for the high-dimensional regime to ensure concentration of the resultant high-dimensional linear models after projection.

4.3.2 Main Results: Independent Sampling Times

We start by considering the low-dimensional regime.

Proposition 4.4. *Consider the model given in equation* (4.3.1) *with* $s_\beta^* = p < n$. *Let* $\mathbf{M} = \mathrm{diag}((m_1, \ldots, m_n))$. *Under assumptions (4.9), (4.14), and (4.15), the MISE is given by*

$$\mathrm{MISE}(\hat{\boldsymbol{\beta}}^{LD}) = \mathrm{tr}\left(\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\left(\mathcal{O}\left(1\right)\mathbf{I}_n + \mathcal{O}(g(n)K_\beta)\mathbf{M}^{-1}\right)\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\right) + \mathcal{O}(s_\beta^* K_\beta^{-2\alpha}).$$

Then, K_β is chosen to minimize the right hand side. In general, the optimal value of K_β is dependent on the specific sequence of designs. We consider a special case where we can characterize the exact tradeoff between the number of unique individuals, n, and the number of samples per observation m_i.

Example 6 (Minimax Estimation of the Mean Function)**.** Consider the model given in equation (4.1.2) with independent uniform sampling times, which is reproduced below for convenience.

$$y_i(t_{i,j}) = \beta^*(t_{i,j}) + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

Here, $s_\beta^* = p = 1$ with $x_i = 1$ for $i = 1, \ldots, n$. In this case, we may set $g(n) = 1$. For now, we write $m = \left(\sum_{i=1}^n m_i^{-1}\right)^{-1}$ to denote the harmonic mean of the $(m_i)_{i=1}^n$. Then,

$$\mathrm{tr}\left(\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{I}_n\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\right) = \mathcal{O}\left(n^{-1}\right).$$

Next, by a direct calculation,

$$\sum_{k=1}^{K_\beta} \operatorname{tr}\left(\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{M}^{-1}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\right) = \mathcal{O}\left(K_\beta(mn)^{-1}\right)$$

Thus, the risk from Proposition 4.4 can be simplified to

$$\operatorname{MISE}(\hat{\boldsymbol{\beta}}^{\mathrm{LD}}) = \mathbb{E}\int_0^1 \left\|\hat{\boldsymbol{\beta}}^{\mathrm{LD}}(t) - \boldsymbol{\beta}^*(t)\right\|^2 dt = \mathcal{O}\left(n^{-1} + K_\beta(mn)^{-1} + K_\beta^{-2\alpha}\right).$$

This yields the optimal choice of $K_\beta \asymp (mn)^{1/(2\alpha+1)}$. Then, the risk is

$$\mathbb{E}\int_0^1 \left\|\hat{\boldsymbol{\beta}}^{\mathrm{LD}}(t) - \boldsymbol{\beta}^*(t)\right\|^2 dt = \mathcal{O}\left(n^{-1} + (mn)^{-2\alpha/(2\alpha+1)}\right),$$

which coincides with the minimax rate from Theorem 3.1 of Cai and Yuan (2011).

The next result is the analogue of Proposition 4.4 for the high-dimensional setting.

Theorem 4.5. *Consider the model given in equation* (4.3.1). *Assume (4.6), (4.9), (4.11), (4.14), and (4.16). For $k \le K_\beta$ and $t > 0$, let*

$$\lambda_{0,k} \triangleq \sqrt{\max_{j=1,\ldots,p} n^{-1}\sum_{i=1}^n \varsigma_{\zeta,i,k}^2 \mathbf{x}_{i,j}^2}\sqrt{\frac{t^2 + 2\log(p)}{n}},$$

where $\varsigma_{\zeta,i,k}^2 = \mathcal{O}(\varsigma_{\mathfrak{o},k}^2 + g(n)m_i^{-1})$. Suppose $\lambda_k \ge 2\lambda_{0,k}$.

1. *If in addition (4.12) holds, then with probability at least $1 - 2\exp(-t^2/2)$,*

$$n^{-1}\left\|\mathbf{X}\left(\hat{\beth}_k - \beth_k^*\right)\right\|_2^2 + \lambda_k\left\|\hat{\beth}_k - \beth_k^*\right\|_1 \le 4\lambda_k^2 s_\beta^*/\phi_{cc,X}^2.$$

2. *If in addition (4.13) holds, then with probability at least $1 - 2\exp(-t^2/2)$,*

$$\left\|\hat{\beth}_k - \beth_k^*\right\|_2^2 = \mathcal{O}\left(\lambda_k^2 s_\beta^* \phi_{adap,X}^{-4}\right).$$

Example 7. As a special case, consider the setting where $m_i = m$ for all $i = 1, \ldots, n$. Then,

$$\lambda_{0,k}^2 = \mathcal{O}\left(\left(\varsigma_{\mathfrak{o},k}^2 + g(n)m^{-1}\right)\frac{\log(p)}{n}\right).$$

Thus, under assumptions (4.9), (4.11), (4.13), (4.14), and (4.16), we have for $k = 1, \ldots, K_\beta$ that

$$\left\|\hat{\beth}_k - \beth_k^*\right\|_2^2 = \mathcal{O}\left(\left(\varsigma_{\mathfrak{o},k}^2 + g(n)m^{-1}\right)\frac{s_\beta^* \log(p)}{n}\right).$$

Then, with probability at least $1 - 2\exp(-t^2/2 + \log(K_\beta))$,

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) = \mathcal{O}\left(\left(1 + K_\beta m^{-1} g(n)\right)\frac{s_\beta^* \log(p)}{n} + s_\beta^* K_\beta^{-2\alpha}\right).$$

This yields an optimal choice of $K_\beta \asymp (nm/(g(n)\log(p)))^{1/(2\alpha+1)}$ with risk

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) = \mathcal{O}\left(\frac{s_\beta^* \log(p)}{n} + s_\beta^* \left(\frac{g(n)\log(p)}{nm}\right)^{2\alpha/(2\alpha+1)}\right).$$

If we assume that $m = 1$, then this corresponds to the model in equation (4.1.3), which is a simplification of the model of Klopp and Pensky (2015) under uniform smoothness. Then, with probability at least $1 - 2\exp(-t^2/2 + \log(K_\beta))$,

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) = \mathcal{O}\left(\frac{s_\beta^* \log(p)}{n} + s_\beta^* \left(\frac{g(n)\log(p)}{n}\right)^{2\alpha/(2\alpha+1)}\right),$$

which, up to the slowly varying functions $g(n)$ and $\log(p)$, achieves the minimax rate from Theorem 1 of Klopp and Pensky (2015).

4.3.3 Main Results: Common Sampling Times

Again, we start by considering the low-dimensional estimator.

Proposition 4.6. *Consider the model given in equation* (4.3.1) *with $s_\beta^* = p < n$. Under assumptions (4.5), (4.10), (4.14), and (4.15), the MISE for $K_\beta \leq m - 1$ is given by*

$$\mathrm{MISE}(\hat{\boldsymbol{\beta}}^{LD}) = \mathcal{O}\left(s_\beta^* m^{-2\alpha} + (1 + K_\beta m^{-1})\mathrm{tr}[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}] + s_\beta^* K_\beta^{-2\alpha}\right).$$

Remark. In Proposition 4.6, if we make a scaling assumption analogous to (4.2) with $\mathrm{tr}[(X^\mathsf{T}X)^{-1}] = \mathcal{O}(s_\beta^*/n)$, then by choosing $K_\beta \leq m - 1$ and $K_\beta \asymp m$, we may further obtain the bound

$$\mathrm{MISE}(\hat{\boldsymbol{\beta}}^{\mathrm{LD}}) = \mathcal{O}\left(s_\beta^* n^{-1} + s_\beta^* m^{-2\alpha}\right).$$

Example 8 (Minimax Estimation of the Mean Function)**.** Consider again the model given in equation (4.1.2) with common sampling times, which is reproduced below for convenience.

$$y_i(t_{i,j}) = \beta^*(t_{i,j}) + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

In this case, it is clear that $g(n) = 1$ satisfies assumption (4.14). Next, observe that

$$\operatorname{tr}\left(\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\right) = n^{-1}.$$

Thus, the MISE simplifies to

$$\mathrm{MISE} = \mathcal{O}\left(n^{-1} + (nm)^{-1}K_\beta + m^{-2\alpha} + K_\beta^{-2\alpha}\right).$$

Choosing $K_\beta \leq m-1$ and $K_\beta \asymp m$, it follows that

$$\mathrm{MISE} = \mathcal{O}\left(n^{-1} + m^{-2\alpha}\right),$$

which coincides with the minimax rate from Theorem 2.1 of Cai and Yuan (2011).

The next result is the analogue of Theorem 4.5 for the common sampling times setting.

Theorem 4.7. *Consider the model given in equation* (4.3.3)*. Assume (4.6), (4.9), (4.11), (4.14), and (4.16). For $k \leq K_\beta$ and $t > 0$, let*

$$c_k \triangleq \begin{cases} \varsigma_{\mathfrak{o},k}^2 + \sqrt{2}\sum_{r=1}^\infty \varsigma_{\mathfrak{o},2rm}^2, & k = 1, \\ \varsigma_{\mathfrak{o},k}^2 + \sum_{r=1}^\infty \varsigma_{\mathfrak{o},2rm+k}^2 + \varsigma_{\mathfrak{o},2rm-k}^2, & k = 2, 4, \ldots, m-1, \\ \varsigma_{\mathfrak{o},k}^2 + \sum_{r=1}^\infty \varsigma_{\mathfrak{o},2rm+k}^2 + \varsigma_{\mathfrak{o},2rm+2-k}^2, & k = 3, 5, \ldots, m-1, \end{cases}$$
$$\lambda_{0,k} \triangleq \sqrt{c_k + m^{-1}\varsigma_\varepsilon^2}\sqrt{\frac{t^2 + 2\log(p)}{n}}.$$

Suppose $\lambda_k \geq 2\lambda_{0,k}$.

1. *If, in addition (4.12) holds, then with probability at least $1 - 2\exp(-t^2/2)$,*

$$\|\mathbf{X}(\hat{\beth}_k^{HD} - \beth_k^* - \daleth_k)\|_2^2 + \lambda_k\|\hat{\beth}_k^{HD} - \beth_k^* - \daleth_k\|_1 \leq 4\lambda_k^2 s_\beta^*/\phi_{cc,X}^2.$$

2. *If, in addition (4.13) holds, then with probability at least $1 - 2\exp(-t^2/2)$,*

$$\|\hat{\beth}_k^{HD} - \beth_k^* - \daleth_k\|_2^2 = \mathcal{O}(s_\beta^*\lambda_k^2).$$

3. *If, in addition (4.13) holds,* $K_\beta \le m-1$, $K_\beta \asymp m$, *and* $\lambda_k = 2\lambda_{0,k}$, *then with probability at least* $1-2\exp(-t^2/2+\log(K_\beta))$,

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{HD}) = \mathcal{O}\left(\frac{s_\beta^* \log(p)}{n} + s_\beta^* m^{-2\alpha}\right).$$

4.4 Two-Stage Estimation

So far, we have only considered the special cases where either $p=0$ or $q=0$. In practice, it is more common to have a mixture of both time varying and time invariant covariates. In this section, we briefly describe how to extend the theory developed in the preceding two sections to consider the general model from Section 5.1, which is reproduced below.

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j})\rangle_2 + \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j})\rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

Since $\boldsymbol{\beta}^*(\cdot)$ is smooth and $\mathbf{x}_i$ is time invariant, the product $\langle \mathbf{x}_i, \boldsymbol{\beta}^*(\cdot)\rangle_2$ is smooth. Thus, we may similarly consider the differencing approach from Section 4.2 to simultaneously remove the effect of $\langle \mathbf{x}_i, \beta^*(\cdot)\rangle_2$ and $\xi_i(\cdot)$. Again, for simplicity, we consider the differencing given by $\mathscr{A}_h$. That is, we may likewise form the differenced linear model from equation (4.2.2)

$$\mathbf{v} = \boldsymbol{\Psi}\daleth^* + \boldsymbol{\eta} + \boldsymbol{\Delta}^{(\gamma)},$$

where

$$\Delta_{i,j}^{(\gamma)} = \sum_{k=K_\gamma+1}^{\infty} \langle \varphi_k(t_{i,(j+1)})\mathbf{z}_i(t_{i,(j+1)}) - \varphi_k(t_{i,(j)})\mathbf{z}_i(t_{i,(j)}), \daleth_k^*\rangle_2 \\ + \xi_i(t_{i,(j+1)}) - \xi_i(t_{i,(j)}) + \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,(j+1)}) - \boldsymbol{\beta}^*(t_{i,(j)})\rangle_2.$$

Then, we may use the same estimators for $\boldsymbol{\gamma}^*(\cdot)$ from Section 4.2. To estimate $\boldsymbol{\beta}^*(\cdot)$, we consider the residuals after estimating $\boldsymbol{\gamma}^*(\cdot)$. That is, define

$$\tilde{y}_i(t_{i,j}) \triangleq y_i(t_{i,j}) - \langle \mathbf{z}_i(t_{i,j}), \hat{\boldsymbol{\gamma}}(t_{i,j})\rangle_2.$$

Then, we may again convert these observations to the frequency domain and use the estimators for $\beta^*(\cdot)$ from Section 4.3. Depending on whether the sampling times are random and independent or fixed and common, we let

$$\omega_{i,k} = \langle \mathbf{x}_i, \beth_k^*\rangle_2 + \zeta_{i,k} + \Delta_{i,k}^{(\beta^*)},$$

or

$$\omega_{i,k} = \langle \mathbf{x}_i, \beth_k^* + \daleth_k \rangle_2 + \zeta_{i,k} + \Delta_{i,k}^{(\beta^*)},$$

where

$$\Delta_{i,k}^{(\beta^*)} \triangleq m_i^{-1} \sum_{j=1}^{m_i} \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j}) - \hat{\gamma}(t_{i,j}) \rangle_2 \varphi_k(t_{i,j}).$$

The arguments are nearly identical to the preceding sections, with the only change being the change in the bias terms. Hence, we omit the results of this section.

4.5 Confidence Bands

In this section, we consider the problem of constructing a confidence band for a particular varying coefficient. The problem of inference in high-dimensional varying coefficient models was first considered by Chen and He (2018). However, they only considered the problem of conducting inference at a prespecified time. To the best of our knowledge, there have been no proposals for confidence bands in high-dimensions. Since the proof technique is similar between the time invariant covariates and the time varying covariates, we only provide the details for $\beta_1^*(\cdot)$ with independent sampling times, but describe the procedure for $\gamma_1^*(\cdot)$. For simplicity, we assume that $m_i = m$ for all $i = 1, \ldots, n$. Let $\sigma_{\zeta,k}^2 \triangleq \mathrm{Var}(\zeta_{i,k})$.

Recall that $\boldsymbol{\beta}_1^*(\cdot)$ admits a decomposition as

$$\boldsymbol{\beta}_1^*(\cdot) = \underline{\boldsymbol{\beta}^*}_1(\cdot) + \overline{\boldsymbol{\beta}^*}_1(\cdot),$$

where $\underline{\boldsymbol{\beta}^*}_1(\cdot)$ and $\overline{\boldsymbol{\beta}^*}_1(\cdot)$ are the low and high frequency components of $\boldsymbol{\beta}_1^*(\cdot)$ respectively. Under some regularity conditions, we may bound the high frequency signal by

$$\left| \sum_{k=K_\beta+1}^{\infty} \beth_{k,1}^* \varphi_k(t) \right| = \mathcal{O}\left(K_\beta^{-\alpha} \log(K_\beta) \right)$$

uniformly for all $t \in (0,1)$. Therefore, it suffices to construct a confidence band for $\boldsymbol{\beta}_1^*(\cdot)$ and tune K_β to balance with the above bias. By the definition of $\underline{\boldsymbol{\beta}^*}_1(\cdot)$, we have that

$$\underline{\boldsymbol{\beta}^*}_1(\cdot) = \sum_{k=1}^{K_\beta} \beth_{k,1}^* \varphi_k(\cdot).$$

For each $k = 1, \ldots, K_\beta$, a confidence interval for $\daleth^*_{k,1}$ may be constructed using the debiased lasso. We write $\hat{\boldsymbol{\Sigma}} \triangleq \mathbf{X}^\intercal \mathbf{X}/n$ and define $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^{p \times p}$ as the relaxed inverse of $\hat{\boldsymbol{\Sigma}}$ using nodewise lasso as in van de Geer et al. (2014). The debiased estimator for $\daleth^*_k$ is given by

$$\hat{\daleth}^{\mathrm{DB}}_k \triangleq \hat{\daleth}^{\mathrm{HD}}_k + \hat{\boldsymbol{\Theta}} \mathbf{X}^\intercal \left(\boldsymbol{\Omega} - \mathbf{X} \hat{\daleth}^{\mathrm{HD}}_k \right) / n.$$

Writing $\hat{\boldsymbol{\theta}}^\intercal$ to denote the first row of $\hat{\boldsymbol{\Theta}}$, the debiased estimator for $\daleth^*_{k,1}$ is

$$\hat{\daleth}^{\mathrm{DB}}_{k,1} \triangleq \hat{\daleth}^{\mathrm{HD}}_{k,1} + \hat{\boldsymbol{\theta}}^\intercal \mathbf{X}^\intercal \left(\boldsymbol{\Omega} - \mathbf{X} \hat{\daleth}^{\mathrm{HD}}_k \right) / n.$$

By using a multiple comparisons correction procedure, we may construct simultaneous confidence intervals for $\daleth^*_{k,1}$ for $k = 1, \ldots, K_\beta$. We may then use these simultaneous confidence intervals to extend to a confidence band for $\underline{\boldsymbol{\beta}^*}(\cdot)$. Let a_k and b_k be the lower and upper bounds for a $1 - \tau$ simultaneous confidence interval of $\daleth^*_{k,1}$. Then, the $1 - \tau$ lower and upper confidence bands for $\underline{\boldsymbol{\beta}^*}_1(\cdot)$ will be given by

$$l^{(\beta^*)}(t) \triangleq \min_{c_k : \{a_k \le c_k \le b_k\}} \sum_{k=1}^{K_\beta} c_k \varphi_k(t), \qquad u^{(\beta^*)}(t) \triangleq \max_{c_k : \{a_k \le c_k \le b_k\}} \sum_{k=1}^{K_\beta} c_k \varphi_k(t).$$

By slightly enlarging these values, we may account for the bias of $\overline{\boldsymbol{\beta}^*}(\cdot)$ asymptotically. That is, for a value of $\delta > 0$, define

$$l^{(\beta^*)}_\delta(t) \triangleq l(t) - \delta, \qquad u^{(\beta^*)}_\delta(t) \triangleq u(t) + \delta.$$

For $\gamma_1(\cdot)$, we may use a similar idea. First, consider the simpler problem of constructing confidence intervals at an arbitrary point $t^* \in (0, 1)$. Let $(t^*) = (\varphi_1(t^*), \ldots, \varphi_{K_\gamma}(t^*))^\intercal$ denote a loading vector, which we identify (t^*) as a vector in $\mathbb{R}^{K_\gamma}$ as well as a vector in $\mathbb{R}^{qK_\gamma}$. Since, as a vector in $\mathbb{R}^{qK_\gamma}$, (t^*) is a sparse loading vector, to construct a confidence interval interval for $\underline{\boldsymbol{\gamma}^*}(\cdot)$, we may consider the approach of Cai and Guo (2017) for estimating linear functionals. This yields a confidence interval for $\underline{\boldsymbol{\gamma}^*}(t^*)$.

Then, consider the time grid $1/(2K_\gamma), 2/(2K_\gamma), \ldots, 2K_\gamma/(2K_\gamma)$. By using a multiple comparisons adjustment, we may construct simultaneous $1-\tau$ confidence intervals at the $2K_\gamma$ time points.

To extend this to a confidence band for $\underline{\boldsymbol{\gamma}^*}(\cdot)$, consider the following lower and upper bounds:

$$l^{(\gamma)}(t) \triangleq \min_{c_k:\{a_k \leq c_k \leq b_k\}} \sum_{k=1}^{K_\gamma} c_k \frac{\sin(\pi(2K_\gamma t - k))}{\pi(2K_\gamma t - k)},$$
$$u^{(\gamma)}(t) \triangleq \max_{c_k:\{a_k \leq c_k \leq b_k\}} \sum_{k=1}^{K_\gamma} c_k \frac{\sin(\pi(2K_\gamma t - k))}{\pi(2K_\gamma t - k)}.$$

At first glance, this definition may seem strange. However, our confidence band is leveraging the Nyquist-Shannon Theorem (cf. Shannon (1949)). Since every low-frequency signal, with maximal frequency K_γ, can be recovered by interpolation of the signal at the grid points $1/(2K_\gamma), 2/(2K_\gamma), \ldots, 2K_\gamma/(2K_\gamma)$ using the sinc function, the above band simultaneously covers all possible interpolations that can arise given the confidence intervals for the signal value. Then, to incorporate the bias in $\overline{\boldsymbol{\gamma}^*}(\cdot)$, we may again enlarge these intervals by a value of $\delta > 0$.

We note that the idea of using a multiple comparisons correction to construct confidence bands is not new in the literature. Earlier works such as Knafl et al. (1985) and Wu et al. (1998) use a Bonferroni adjustment at various gridpoints and interpolate over the interval by bounding the derivative. Conversely, for $\boldsymbol{\beta}^*(\cdot)$, we construct simultaneous confidence intervals on the Fourier coefficients, which induces simultaneous confidence intervals for all linear combinations. Likewise, for $\boldsymbol{\gamma}^*(\cdot)$, instead of bounding the derivatives, we exploit the Nyquist-Shannon Theorem to provide uniformity over the entire interval.

4.5.1 Assumptions

(4.17) The sparsity s_β^* satisfies $s_\beta^* = o(\sqrt{n}/\log(p))$.

(4.18) The projected error term $\boldsymbol{\zeta}$ satisfies $\sigma_{\zeta,k}^2 \triangleq \text{Var}(\zeta_{i,k}) \asymp \varsigma_{\zeta,k}^2$.

(4.19) The estimator $\hat{\boldsymbol{\Theta}}$ satisfies $\sqrt{n}\|\mathbf{I}_p - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log(p)})$.

Remark. The first assumption (4.17) is the standard sparsity requirement in high-dimensional inference. Next, (4.18) is a technical requirement that is satisfied, for example, by the Gaussian distribution. Sufficient conditions for (4.19) are given in van de Geer et al. (2014).

4.5.2 Main Results

Theorem 4.8. *Consider the model given in equation* (4.3.1). *Assume (4.5), (4.9), (4.14), (4.16), and (4.17) – (4.19). Moreover, let the tuning parameters for lasso satisfy* $\lambda_k \asymp \lambda_{0,k}$ *and* $\lambda_k \geq 3\lambda_{0,k}$

for $k = 1, \ldots, K_\beta$. Similarly, letting ν_j for $j = 1, \ldots, p$ denote the tuning parameters for nodewise lasso, assume that the parameters satisfy $\nu_j \asymp K_0\sqrt{\log(p)/n}$ and $\nu_j \geq 3K_0\sqrt{\log(p)/n}$.

For each $k = 1, \ldots, K_\beta$, write

$$\begin{aligned}\sqrt{n\sigma_{\zeta,k}^{-2}}\left(\hat{\beth}_{k,1}^{DB} - \beth_{k,1}^{*}\right) &= W_k + \Delta_k,\\ W_k &= \sigma_{\zeta,k}^{-1}\hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\zeta_k/\sqrt{n},\\ \Delta_k &= \sqrt{n\sigma_{\zeta,k}^{-2}}\left(\left(\mathbf{I}_p - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\right)\left(\hat{\beth}_k^{HD} - \beth_k^{*}\right)\right)_1.\end{aligned}$$

Let $(V_1, \ldots, V_{K_\beta})^{\mathsf{T}} \sim \mathcal{N}_{K_\beta}(\mathbf{0}_{K_\beta}, Var((W_1, \ldots, W_{K_\beta})^{\mathsf{T}}|X))$.

1. *Letting $\mathcal{A}$ denote the class of all hyperrectangles in $\mathbb{R}^{K_\beta}$, then*

$$\sup_{A\in\mathcal{A}}\left|\mathbb{P}\left((W_1, \ldots, W_{K_\beta})^{\mathsf{T}} \in A|\mathbf{X}\right) - \mathbb{P}\left((V_1, \ldots, V_{K_\beta})^{\mathsf{T}} \in A|\mathbf{X}\right)\right| \to 0.$$

2. *Then,*

$$\sup_{k=1,\ldots,K_\beta}|\Delta_k| = o_{\mathbb{P}}(1).$$

Remark. In practice, one may use the scaled lasso of Sun and Zhang (2012) to estimate $\sigma_{\zeta,k}^2$, denoted by $\hat{\sigma}_{\zeta,k}^2$.

For concreteness, we consider simultaneous confidence intervals constructed by Bonferroni correction. Let z_τ denote the τ upper quantile of a standard Gaussian distribution. Then, the confidence intervals for $\beth_{k,1}^*$ are given by $\hat{\beth}_{k,1}^{\mathrm{DB}} \pm z_{\tau/(2K_\beta)}\hat{\sigma}_{\zeta,k}/\sqrt{n}$. Then, we have the following proposition regarding the coverage and asymptotic performance of the confidence band.

Proposition 4.9. *Consider the setup of Theorem 4.8 with the simultaneous confidence intervals constructed by Bonferroni correction. Suppose that $K_\beta \asymp \min((n\log(n))^{1/(2\alpha)}, (nm\log(n)/g(n))^{1/(2\alpha+2)})$ and $\delta \asymp K_\beta^{-\alpha}\log(K_\beta)$.*

1. *For n sufficiently large,*

$$\mathbb{P}\left(\forall t \in (0,1) : l_\delta^{(\beta^*)}(t) \leq \boldsymbol{\beta}_1^*(t) \leq u_\delta^{(\beta^*)}(t)\right) \geq 1 - \tau.$$

2. *For all $t \in (0,1)$,*

$$\left|u_\delta^{(\beta^*)}(t) - l_\delta^{(\beta^*)}(t)\right| \leq \max\left((n\log(n))^{-1/2}, (nm\log(n)/g(n))^{-\alpha/(2\alpha+2)}\log(n)\right).$$

Remark. We note that the maximal width of the confidence band has two rates, depending on the growth rate of m relative to n. If m is very large, then most of the error in the resultant linear model comes from the random Fourier coefficients of $\xi_i(\cdot)$. Since the variance for the high-frequency components is relatively low, this encourages oversmoothing to reduce the bias. This leads to a near parametric rate of $\sqrt{\log(n)/n}$ in the width. Conversely, when m is small, the noise is dominated by the inexact orthogonalization. This variance accumulates as we increase the number of frequencies to be estimated, resulting in less smoothing compared to the large m setting.

This phenomenon is related to the two part rate for estimation, as seen in Example 7. After a certain threshold of m, additional observations per subject do not improve the risk in estimation since the bottleneck is in averaging the random effects.

4.6 Simulations

The general model that we consider is given by

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j})\rangle_2 + \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j})\rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}),$$

Throughout, the covariates and the noise are independent and identically distributed standard Gaussian variables; that is, $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$, $\mathbf{z}_i(t_{i,j}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\mathbf{0}_q, \mathbf{I}_q)$, and $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. When generating the varying coefficients and the random effects, we use either the trigonometric basis or by the B-spline basis; the choice of the coefficients is described below. We consider both fixed and common sampling times as well as random and independent sampling times. When the sampling times are fixed and common (denoted "com"), we set $t_{i,j} = j/m$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$. On the other hand, when the sampling times are random and independent (denoted "ind"), we let $t_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$.

All of our tuning parameters, λ, K_β, and K_γ are chosen via five-fold cross-validation. For each combination of parameter values that we consider, we simulate the data 200 times and compute the high-dimensional estimator. To evaluate the performance of our estimator, we consider three metrics: average loss, average coverage, and average length. Average loss is integrated squared error in estimating $\boldsymbol{\beta}^*(\cdot)$ averaged over the trials. Average coverage is the proportion of times the confidence band covers the true varying coefficient function $\boldsymbol{\beta}_1^*(\cdot)$ with a nominal coverage of 95% and average length is $\max_{t\in(0,1)}(u^{(\beta^*)}(t) - l^{(\beta^*)}(t))$ averaged over the trials. We consider average loss for both $\boldsymbol{\gamma}^*(\cdot)$ and $\boldsymbol{\beta}^*(\cdot)$ while average coverage and average length are only considered for $\boldsymbol{\beta}^*(\cdot)$. For our simulations, we consider the two special cases of Section 4.2 and 4.3.

In the setting when $q = 0$, we have

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

When $\boldsymbol{\beta}^*(\cdot)$ is generated from the trigonometric basis, the Fourier coefficients are given by

$$\beth^*_{k,l} = \begin{cases} \zeta_{k,l}((k+1)/2)^{-2.1} & k = 1, 3, \ldots, 29 \text{ and } l = 1, \ldots, s^*_\beta \\ \zeta_{k,l}((k+2)/2)^{-2.1} & k = 2, 4, \ldots, 30 \text{ and } l = 1, \ldots, s^*_\beta \\ 0 & \text{else} \end{cases}$$

where $\zeta_{k,l} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1)$. Then, we rescale $\beth^*_k$ such that $\int_0^1 (x_i^\intercal \beta^*(t))^2 dt = 4$ to keep the signal to noise ratio constant. For the B-spline basis, we consider

$$\beth^*_{k,l} = \begin{cases} \zeta_{k,l} & k = 1, 2, 3 \text{ and } l = 1, \ldots, s^*_\beta \\ 0 & \text{else,} \end{cases}$$

which is then rescaled to keep the signal to noise ratio constant. Similarly, for the random effects, $\xi_i(\cdot)$, we rescale the coefficients so the variance is constant at one.

In this setting, we let $n \in \{200, 500\}$, $m \in \{25, 50, 75, 150\}$, $p \in \{500, 1000\}$, and $s^*_\beta \in \{15, 25\}$. The results are presented in Tables A.4.1 and A.4.2 in Section A.4.6 of the Supplement for the trigonometric and B-spline basis respectively. For both bases, consistent with Theorems 4.5 and 4.7, as s^*_β increases, the average loss increases while the loss decreases as n or m increase. Surprisingly, the loss for the fixed and common sampling times seems to be better than for the random and independent sampling times for the same value of n and m despite the rate of convergence for the random and independent sampling times being faster. Similarly, the confidence bands exhibit higher coverage with shorter lengths in the fixed and common sampling times as opposed to the random and independent sampling times. We note that the confidence bands for the spline basis are significantly wider than for the trigonometric basis to account for the fact that the trigonometric basis functions periodic. Since the spline functions are aperiodic, the confidence bands are wider to be able to cover allow coverage at both endpoints. In Figure 4.1, the plot on the left shows a confidence band with the B-spline basis and the plot on the right with the trigonometric basis; note the difference in the widths of the two bands.

In the setting when $p = 0$, we have

$$y_i(t_{i,j}) = \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j}) \rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

Here, we set $n = 200$, $m \in \{25, 50, 75\}$, $q = 500$, and $s^*_\gamma \in \{15, 25\}$. The coefficients $\beth^*_{k,l}$ are

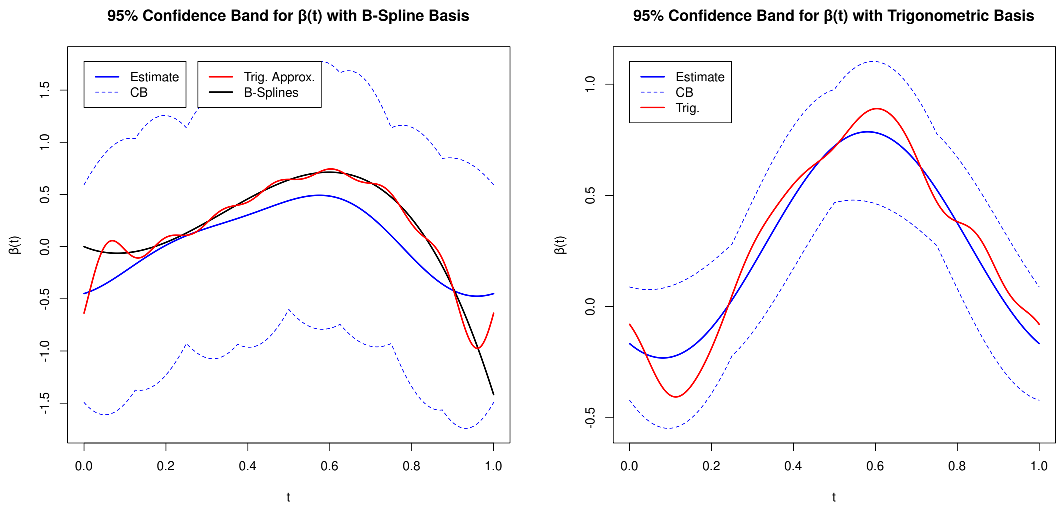

Figure 4.1: 95% Confidence bands for $\boldsymbol{\beta}_1^*(t)$ when $n = 500$, $m = 50$, $s_\beta^* = 15$, and $p = 1000$ with independent and random sampling times. On the left, the data is generated using the b-spline basis. Then, "B-Splines" denotes the true signal, "Trig. Approx." denotes the best approximation of $\boldsymbol{\beta}_1^*(t)$ using 30 trigonometric basis function, and "Estimate" and "CB" are the estimate and confidence band from Section 4.5 respectively. On the right, the data is generated using the trigonometric basis. Then, "Trig." denotes the true signal and "Estimate" and "CB" are the estimate and confidence band from Section 4.5 respectively.

generated similar to $\daleth^*_{k,l}$. The results are presented in Table A.4.3 in Section A.4.6 of the Supplement. We see that $\mathscr{A}_h$ outperforms $\mathscr{B}_h$ when the sampling times are independent and random whereas $\mathscr{B}_h$ outperforms $\mathscr{A}_h$ when the sampling times are common and fixed. As s_γ^* increases, the estimation error increases while the estimation error decreases as m increases, which is consistent with our theoretical results. Moreover, similar to estimating $\boldsymbol{\beta}^*(\cdot)$, our estimation error is higher for the B-spline basis as compared to the trigonometric basis.

4.7 Human Height Data

In this section, we are interested in analyzing the average height across countries. The height data is freely available from the NCD Risk Factor collaboration at https://ncdrisc.org/index.html, which includes the average height of birth cohorts aged five through nineteen over many decades for both sexes. A detailed description of the data is provided in Rodriguez-Martinez et al. (2020). Although the data contains 95% credible intervals for the average height in a country of each age group at a particular time, our response comprises solely of the point estimate. To supplement this data, we use a variety of United Nations data on countries, including the World Health Organization (https://www.who.int/data/collections), the Human

Development Reports (http://www.hdr.undp.org/en/data), and the United Nations International Children's Emergency Fund (https://data.unicef.org/). In addition, we also use information on caloric, protein, and fat supply from Our World in Data (https://ourworldindata.org/food-supply). For covariates that are time varying but not collected annually, we impute the values for intermediate years using an exponentially weighted moving average. We focus on the average height of five year old boys, the youngest age in the dataset. The model that we consider is model (4.1.1), which is reproduced below for convenience:

$$y_i(t_{i,j}) = \langle \mathbf{x}_i, \boldsymbol{\beta}^*(t_{i,j})\rangle_2 + \langle \mathbf{z}_i(t_{i,j}), \boldsymbol{\gamma}^*(t_{i,j})\rangle_2 + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

Since some covariates are only available for a subset of countries, we only consider the $n = 98$ countries that have some measurements for all covariates. We constrain our analysis to the years 1990 to 2019 to limit the imputation of our time-varying covariates. Therefore, we have common sampling times, with $m = 30$. When normalizing our sampling times to the unit interval, $t = 0$ and $t = 29/30$ correspond to 1990 and 2019 respectively. Our time invariant covariates are various classifications of the countries according to the United Nations; in particular, we consider development status and geographic regions and sub-regions, with sub-regions being nested within regions. Expanding out these groupings, we have $p = 25$ time invariant covariates. Our time varying covariates consist of various socioeconomic factors, such as human development index, urbanization rate, gross domestic product per capita, etc. When synthesizing the height data, Finucane et al. (2014) considered the interaction between income and urbanization rate. Thus, we include all possible two-way interactions of our time-varying covariates, leaving us with $q = 212$ covariates. Finally, the random effects represent the unobserved heterogeneity amongst countries, which, as mentioned in the Introduction, encapsulates environmental factors. These environmental factors are potentially correlated with our observed time-varying covariates.

The goal in our data analysis is to analyze the trend in the average height, controlling for other covariates. That is, we are interested in estimating the intercept term. Since our time-invariant covariates consist of many geographic regions, our reference region is North Africa. A plot of the estimate is given in Figure 4.2.

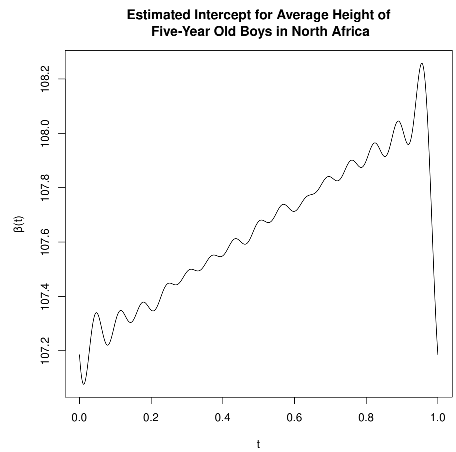

Figure 4.2: Estimated coefficient for the intercept, with North Africa as the reference region.

From the plot, we notice that the estimate of the average height in North Africa, controlling for other covariates, is almost linearly increasing, which is not specified apriori. Note that the precipitous drop near $t = 1$ is an artifact of the estimation procedure using the trigonometric basis. As we do not believe that the average height over the span of thirty years is periodic, our estimator $\hat{\boldsymbol{\beta}}(\cdot)$ is estimating the best periodic approximation to the underlying non-periodic intercept function. Thus, the interpretation of the plot is valid only away from the two endpoints.

CHAPTER 5

Rank-Constrained Least-Squares: Prediction and Inference

5.1 Introduction

In this work, we focus on the trace regression model:

$$y = \langle \mathbf{X}, \boldsymbol{\Theta}^* \rangle_{\mathrm{HS}} + \varepsilon. \tag{5.1.1}$$

Here y is a real-valued response, $\mathbf{X}$ is a feature matrix valued in $\mathbb{R}^{d_1 \times d_2}$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is the parameter of our interest, and ε is a noise term that is independent of $\mathbf{X}$. Throughout, objects with a superscript * denote true model paramters and we define $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ as the trace inner product in the sense that given any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$, $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{HS}} \triangleq \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. Throughout, we write $d \triangleq d_2$ and assume without the loss of generality that $d_1 \le d_2$ by possibly transposing the data.

Suppose we have n independent observations $(\mathbf{X}_i, y_i)_{i \in [n]}$ generated from model (5.1.1). Under a high-dimensional setup, n is much smaller than $d_1 d_2$, and some structural assumptions on $\boldsymbol{\Theta}^*$ are necessary to reduce the degrees of freedom of $\boldsymbol{\Theta}^*$ to achieve estimation consistency. Here, we assume that $\boldsymbol{\Theta}^*$ is low-rank; that is, $r^* \triangleq \mathrm{rank}(\boldsymbol{\Theta}^*)$ with $r^* d_1 \ll n$. Given that one needs at most $(2r^* + 1)d_1$ parameters to determine $\boldsymbol{\Theta}^*$ through a singular value decomposition, intuitively a sample of size n should suffice to achieve estimation consistency. The high-dimensional low-rank trace regression model was first introduced by Rohde and Tsybakov (2011) and admits many special cases of wide interest. For instance, when $\boldsymbol{\Theta}^*$ and $\mathbf{X}$ are diagonal, model (5.1.1) reduces to a sparse linear regression model:

$$y = \langle \mathbf{x}, \boldsymbol{\beta}^* \rangle_2 + \varepsilon, \tag{5.1.2}$$

where $\boldsymbol{\beta}^* = \mathrm{diag}(\boldsymbol{\Theta}^*)$. Note that $\boldsymbol{\beta}^*$ is sparse because $\|\boldsymbol{\beta}^*\|_0 = r^* \ll d_1$. When $\mathbf{X}$ is a singleton in the sense that $\mathbf{X} = \mathbf{e}_i \mathbf{e}_j^\top$, where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the ith and jth canonical basis vectors respectively, model (5.1.1) reduces to a low-rank matrix completion problem (Candès and Recht (2009), Koltchinskii et al. (2011), Recht (2011), Negahban and Wainwright (2012)).

Perhaps the most natural approach to incorporate the low-rank structure in estimating $\mathbf{\Theta}^*$ is to enforce a rank-constraint directly. Consider the following rank-constrained least-squares estimator:

$$\hat{\mathbf{\Theta}}_{L_0}(r) = \underset{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2},\mathrm{rank}(\mathbf{\Theta})\le r}{\arg\min} \sum_{i=1}^{n} \left(y_i - \langle \mathbf{X}_i, \mathbf{\Theta}\rangle_{\mathrm{HS}}\right)^2. \tag{5.1.3}$$

Note that the rank constraint is non-convex, thereby imposing a fundamental challenge computationally in obtaining this estimator. To resolve this issue, one can resort to nuclear-norm regularization to encourage low-rank structure of the estimator. Specifically, for some $\lambda > 0$, consider

$$\hat{\mathbf{\Theta}}_{\mathrm{N}}(\lambda) = \underset{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2}}{\arg\min} \left\{\frac{1}{2n}\sum_{i=1}^{n} \left(y_i - \langle \mathbf{X}_i, \mathbf{\Theta}\rangle_{\mathrm{HS}}\right)^2 + \lambda\|\mathbf{\Theta}\|_{\mathrm{N}}\right\}, \tag{5.1.4}$$

where $\|\cdot\|_{\mathrm{N}}$ denotes the nuclear norm. Problem (5.1.4) is convex and thus amenable to polynomial-time algorithms. The past decade or so has witnessed a flurry of works on statistical guarantees for $\hat{\mathbf{\Theta}}_{\mathrm{N}}$; a partial list includes Negahban and Wainwright (2011), Rohde and Tsybakov (2011), Candès and Plan (2011), and Fan et al. (2021), among others. For instance, with a restricted strong convexity assumption on the loss function, Negahban and Wainwright (2011) showed that with an appropriate choice of λ, $\|\hat{\mathbf{\Theta}}_{\mathrm{N}}(\lambda) - \mathbf{\Theta}^*\|_{\mathrm{F}}$ is of the order $\sqrt{rd_1/(\kappa n)}$ up to a logarithmic factor, where κ is a lower bound of the minimum restricted eigenvalue (Bickel et al. (2009)) of the Hessian matrix of the loss function.

To the best of our knowledge, it remains open whether κ is inevitable for statistical guarantees on learning $\mathbf{\Theta}^*$. At this point, it is instructive to recall related results for sparse high-dimensional linear regression. Zhang et al. (2014) showed that under a standard conjecture in computational complexity, the in-sample mean-squared prediction error of any estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{poly}}$, that can be computed within polynomial time has the following worst-case lower bound:

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{poly}} - \boldsymbol{\beta}^*\rangle_2^2\right\} \gtrsim \frac{(r^*)^{1-\delta}\log d_1}{n\kappa}, \tag{5.1.5}$$

where δ is an arbitrarily small positive scalar. This result demonstrates the indispensable dependence on κ for any polynomial-time estimator of $\boldsymbol{\beta}^*$, which includes convex estimators like lasso. On the other hand, Bunea et al. (2007) and Raskutti et al. (2011) showed that the L_0-constrained estimator $\hat{\boldsymbol{\beta}}_{L_0}$ (also known as the best subset selection estimator), which is defined as

$$\hat{\boldsymbol{\beta}}_{L_0}(r) \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_1},\|\boldsymbol{\beta}\|_0\leq r}{\arg\min} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle_2)^2, \tag{5.1.6}$$

satisfies the following κ-free prediction error bound:

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}(r^*) - \boldsymbol{\beta}^*\rangle_2^2\right\} \lesssim \frac{r^*\log d_1}{n}. \tag{5.1.7}$$

This demonstrates the robustness of $\hat{\boldsymbol{\beta}}_{L_0}$ against collinearity in the design. However, under the general trace regression model, there are currently no κ-free statistical guarantees for the rank-constrained estimator $\hat{\boldsymbol{\Theta}}_{L_0}$.

The first contribution of our work is an in-sample prediction error bound for the rank-constrained least-squares estimator $\hat{\boldsymbol{\Theta}}_{L_0}$ without a restricted strong convexity requirement. We emphasize that this result is much more challenging to achieve than the counterpart result (5.1.7) for $\hat{\boldsymbol{\beta}}_{L_0}$ and requires a completely different technical treatment. We shall see in the sequel that the in-sample prediction error of both $\hat{\boldsymbol{\Theta}}_{L_0}$ and $\hat{\boldsymbol{\beta}}_{L_0}$ boils down to a supremum process of projections of the noise vector $(\varepsilon_1, \ldots, \varepsilon_n)^\top$ onto a family of low-dimensional subspaces. For $\hat{\boldsymbol{\beta}}_{L_0}$, the family of subspaces is finite; for $\hat{\boldsymbol{\Theta}}_{L_0}$, however, the family of subspaces is a continuous subset of a Stiefel manifold, which is infinite. The main technical challenge we face here is to characterize the complexity of this infinite subspace family. In Theorem 5.2, we leverage a real algebraic geometry tool due to Basu et al. (2007) to bound the Frobenius-norm-based covering number of this family of subspaces.

We then investigate a permutation test for the presence of sparse and low-rank signals respectively as applications of the previous results. In the context of hypotheses testing for high-dimensional sparse linear models, Cai and Guo (2020) and Javanmard and Lee (2020) both consider a debiasing-based test that controls the probability of type-I error uniformly over the null parameter space of sparse vectors. There, the sparsity s^* of the regression coefficients needs to satisfy $s^* = o\{n^{1/2}/\log(p)\}$ for the asymptotic variance of the test statistic to dominate the bias. By considering a permutation test, we circumvent the challenge of characterizing the asymptotic distribution of a test statistic and accommodate denser alternative parameters. Moreover, under a mild assumption on the design, we are able to leverage the super-efficiency of the origin, which was rather seen as a challenge in high-dimensional group inference (Guo et al. (2021)), to test at a faster rate than $n^{-1/2}$. To the best of our knowledge, this is the first proposal to conduct inference for the presence of low-rank signals.

5.1.1 Organization of the Chapter

In Section 5.2, we consider a discretization scheme of all possible models in low-rank trace regression and derive the covering number of the corresponding Stiefel sub-manifold that is used to analyze the performance of $\hat{\mathbf{\Theta}}_{L_0}$ for in-sample prediction. Next, in Section 5.3, we consider global hypotheses testing in signal plus noise models. We start with a general power analysis for signal plus noise models in Section 5.3.1, which we then apply to the sparse high-dimensional linear model and low-rank trace regression model in Sections 5.3.2 and 5.3.3 respectively. By leveraging the projection structure of the rank-constrained estimator, in Section 5.3.4, we demonstrate the robustness of our power analysis to misspecification of the rank. Finally, we analyze the empirical performance of our proposed methodologies in Section 5.4. For the ease of presentation, most of the proofs for Section 5.2 and all of the proofs for Section 5.3 are deferred to Section A.4.6.

5.2 In-Sample Prediction Risk of the Rank-Constrained Estimator

Given an estimator $\hat{\mathbf{\Theta}}$ of $\mathbf{\Theta}^*$, define its in-sample prediction risk as

$$\mathcal{R}(\hat{\mathbf{\Theta}}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2. \tag{5.2.1}$$

This section focuses on characterizing the in-sample prediction risk of the rank-constrained estimator $\hat{\mathbf{\Theta}}_{L_0}$. For any $\mathbf{\Theta}^*$ with rank r^*, there exist two matrices $\mathbf{U}^* \in \mathbb{R}^{d_1\times r^*}$ and $\mathbf{V}^* \in \mathbb{R}^{d_2\times r^*}$ such that $\mathbf{\Theta}^* = \mathbf{U}^*\mathbf{V}^{*\mathsf{T}}$. The existence of $\mathbf{U}^*$ and $\mathbf{V}^*$ is guaranteed, for example, by a singular value decomposition of $\mathbf{\Theta}^*$. Note that this representation is not unique, since for any invertible matrix $\mathbf{A} \in \mathbb{R}^{r\times r}$, we have $\mathbf{\Theta}^* = \mathbf{U}^*\mathbf{A}\mathbf{A}^{-1}\mathbf{V}^{*\top}$. Now the trace regression model (5.1.1) can be represented as

$$y = \langle \mathbf{X}, \mathbf{\Theta}^*\rangle_{\text{HS}} + \varepsilon = \langle \mathbf{X}, \mathbf{U}^*\mathbf{V}^{*\mathsf{T}}\rangle_{\text{HS}} + \varepsilon = \langle \mathbf{X}\mathbf{V}^*, \mathbf{U}^*\rangle_{\text{HS}} + \varepsilon. \tag{5.2.2}$$

Throughout, for a matrix $\mathbf{A} \in \mathbb{R}^{k_1\times k_2}$, we write $\text{vec}(\mathbf{A}) \in \mathbb{R}^{k_1 k_2}$ to denote the vectorization of $\mathbf{A}$. For any $\mathbf{V} \in \mathbb{R}^{d_2\times r}$, let $\mathbf{X}_{\mathbf{V}} \in \mathbb{R}^{n\times rd_1}$ denote the matrix whose ith row is $\text{vec}(\mathbf{X}_i\mathbf{V})$. Writing $\gamma_{\mathbf{U}} \triangleq \text{vec}(\mathbf{U})$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, we then deduce from (5.2.2) that

$$\mathbf{y} = \mathbf{X}_{\mathbf{V}^*}\gamma_{\mathbf{U}} + \boldsymbol{\varepsilon}.$$

Figure 5.1 illustrates the construction of $\mathbf{X}_{\mathbf{V}^*}$ when $r^* = 2$. Suppose $\mathbf{V}^*$ is known in ad-

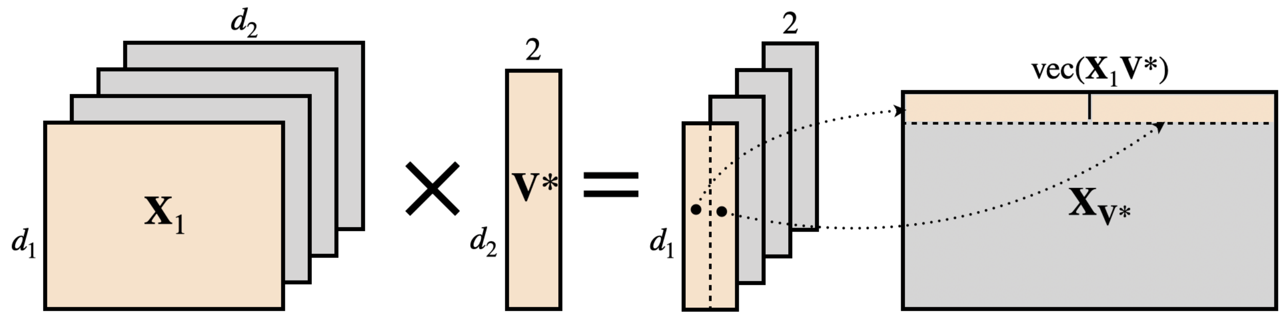

Figure 5.1: Illustration of the construction of $\mathbf{X}_{\mathbf{V}^*}$ with oracle $\mathbf{V}^*$ when $r^* = 2$.

vance. When $n \gg rd_1$, $\gamma_{\mathbf{U}}$ can be consistently estimated by ordinary least-squares to yield an estimator of $\mathbf{\Theta}^*$. Given that ordinary least-squares is projecting $\mathbf{y}$ onto the column space of $\mathbf{X}_{\mathbf{V}}$, the rank-constrained least-squares problem (5.1.3) reduces to finding the optimal $\mathbf{V}$ so that the resulting $\mathbf{X}_{\mathbf{V}}$ captures the most variation of the response $\mathbf{y}$. This motivates our initial step to analyze the in-sample prediction risk. For any $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$, define the projection matrix $\mathbf{P}_{\mathbf{V}} := \mathbf{X}_{\mathbf{V}}(\mathbf{X}_{\mathbf{V}}^\top \mathbf{X}_{\mathbf{V}})^{-1}\mathbf{X}_{\mathbf{V}}^\top$. The following lemma shows that the in-sample prediction risk of $\hat{\mathbf{\Theta}}_{L_0}$ can be bounded by the supremum of projections of the noise vector $\boldsymbol{\varepsilon}$ onto the column space of $\mathbf{X}_{\mathbf{V}}$.

Lemma 5.1. *Consider the model in equation* (5.1.1)*. If $r \geq r^*$, then, the rank-constrained least-squares estimator satisfies*

$$\mathcal{R}(\hat{\mathbf{\Theta}}_{L_0}) \leq \frac{4}{n} \sup_{\mathbf{V} \in \mathbb{R}^{d_2 \times 2r^*}} \|\mathbf{P}_{\mathbf{V}}\boldsymbol{\varepsilon}\|_2^2. \tag{5.2.3}$$

Lemma 5.1 suggests that the complexity of the set $\mathcal{P} \triangleq \{\mathbf{P}_{\mathbf{V}}\}_{\mathbf{V} \in \mathbb{R}^{d_2 \times 2r^*}}$ determines the in-sample prediction risk of $\hat{\mathbf{\Theta}}_{L_0}(r^*)$. Note that the number of columns of $\mathbf{V}$ is $2r^*$ instead of r^*, which is due to the fact that the maximum rank of $(\hat{\mathbf{\Theta}}_{L_0}(r^*) - \mathbf{\Theta}^*)$ is $2r^*$. To quantify this complexity, we consider the metric space $(\mathcal{P}, \|\cdot\|_{\text{HS}})$. We say that $\mathcal{N}_\delta \subseteq \mathcal{P}$ is an δ-net of $\mathcal{P}$ if for any $\mathbf{P} \in \mathcal{P}$, there exists a $\tilde{\mathbf{P}} \in \mathcal{N}_\delta$ such that $\|\mathbf{P} - \tilde{\mathbf{P}}\|_{\text{HS}} \leq \delta$. We define the covering number, $N_\delta(\mathcal{P})$, as the minimum cardinality of an δ-net of $\mathcal{P}$. The following theorem leverages a result from real algebraic geometry to bound $N_\delta(\mathcal{P})$. To the best of our knowledge, this tool is new to statistical analyses in high-dimensions. To highlight the power of the tool, we give the proof immediately after the statement of the theorem.

Theorem 5.2. *For any $\delta < 1$, we have that*

$$N_\delta(\mathcal{P}) \leq 2^{r^* d_1} \left\{ \frac{12 r^* d_1 n^3}{\delta} \right\}^{r^* d_2 + 1}.$$

Proof of Theorem 5.2. To simplify notation, we drop the superscript * in r^* for convenience within this proof. For an integer k, let $[k] \triangleq \{1, \ldots, k\}$. Now, for $\mathcal{S} \subseteq [rd_1]$, write $\mathbf{X}_{\mathbf{V},\mathcal{S}}$ to denote the $\mathbb{R}^{n\times|\mathcal{S}|}$ submatrix of $\mathbf{X}_\mathbf{V}$ with the columns indexed by $\mathcal{S}$. Then, for any fixed $\mathcal{S} \subseteq [rd_1]$, define the collection of projection matrices $\mathcal{P}_\mathcal{S}$ as

$$\mathcal{P}_\mathcal{S} \triangleq \{\mathbf{X}_{\mathbf{V},\mathcal{S}}(\mathbf{X}_{\mathbf{V},\mathcal{S}}^\intercal \mathbf{X}_{\mathbf{V},\mathcal{S}})^{-1}\mathbf{X}_{\mathbf{V},\mathcal{S}}^\intercal : \mathbf{V} \in \mathbb{R}^{d_2\times 2r}, \det(\mathbf{X}_{\mathbf{V},\mathcal{S}}^\intercal \mathbf{X}_{\mathbf{V},\mathcal{S}}) \neq 0\}.$$

Note that

$$\mathcal{P} \triangleq \{\mathbf{P}_\mathbf{V}\}_{\mathbf{V}\in\mathbb{R}^{d_2\times r}} \subseteq \bigcup_{\mathcal{S}\subseteq[rd_1]} \mathcal{P}_\mathcal{S}.$$

To further simplify notation, throughout this proof, we identify the matrix $\mathbf{V} \in \mathbb{R}^{d_2\times r}$ with the vector $\mathbf{v} \in \mathbb{R}^{rd_2}$ by viewing $\mathbf{v}$ as the vectorization of $\mathbf{V}$. Fix $\{\mathbf{X}_i\}_{i\in[n]}$ and $i, j \in [n]$ and consider the map

$$\begin{aligned}\Phi_{ij} : \mathbb{R}^{rd_2}\backslash\{\mathbf{v} : \det(\mathbf{X}_{\mathbf{V},\mathcal{S}}^\intercal \mathbf{X}_{\mathbf{V},\mathcal{S}}) = 0\} &\to [-1, 1],\\ \mathbf{v} \mapsto (\mathbf{P}_{\mathbf{V},\mathcal{S}})_{i,j} &= \{\mathbf{X}_{\mathbf{V},\mathcal{S}}(\mathbf{X}_{\mathbf{V},\mathcal{S}}^\top \mathbf{X}_{\mathbf{V},\mathcal{S}})^{-1}\mathbf{X}_{\mathbf{V},\mathcal{S}}^\top\}_{ij}.\end{aligned}$$

We claim that Φ_{ij} is a rational function of polynomials of order at most $2|\mathcal{S}|$. To see this, for any invertible matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $u, t \in [p]$, the (u, t) entry of the adjugate of $\mathbf{A}$ is given by $\operatorname{adj}(\mathbf{A})_{ut} \triangleq (-1)^{u+t} \det(\mathbf{A}_{-u,-t})$. Then, by Cramer's rule,

$$\mathbf{A}_{i,j}^{-1} = (\det(\mathbf{A}))^{-1} \operatorname{adj}(\mathbf{A})_{ij}.$$

Given that each entry of $\mathbf{X}_\mathbf{V}^\intercal \mathbf{X}_\mathbf{V} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is a quadratic function with respect to $\mathbf{v}$, it follows that $(\det(\mathbf{X}_\mathbf{V}^\intercal \mathbf{X}_\mathbf{V}))^{-1}$ is a polynomial of order at most $2|\mathcal{S}|$ and $\operatorname{adj}(\mathbf{A})_{ut}$ is a polynomial of order at most $2|\mathcal{S}| - 2$. Hence, each entry of $\mathbf{P}_\mathbf{V}$ is a rational function of polynomials of order at most $2|\mathcal{S}|$. Denote the (i, j) entry of $\mathbf{P}_\mathbf{V}$ by $\Phi_{i,j}(\mathbf{v})$, which has representation $\Phi_{i,j}(\mathbf{v}) = F_{i,j}(\mathbf{v})/\gamma(\mathbf{v})$ for polynomials $F_{i,j}(\mathbf{v})$ and $\gamma(\mathbf{v})$ in the domain of $\Phi_{i,j}$.

Now for any $\delta > 0$, consider a monotonically increasing sequence $-1 = s_1 < \ldots < s_m = 1$ such that $m = \lceil 2/\delta \rceil + 1$ and $|s_{t+1} - s_t| \leq \delta$ for any $t \in [m - 1]$. Consider the level sets: $\mathcal{C}_{ijt} \triangleq \{\mathbf{v} \in \mathbb{R}^{2rd_2} : (F_{i,j}(\mathbf{v}) - \gamma(\mathbf{v})s_t) = 0\}, t \in [m]$. Note that these level sets partition the entire $\mathbb{R}^{rd_2}$ into multiple connected components, within each of which any two points $\mathbf{v}_1, \mathbf{v}_2$ satisfy $|\Phi_{ij}(\mathbf{v}_1) - \Phi_{ij}(\mathbf{v}_2)| \leq \delta$. Consider the union of all such level sets over i, j, t:

$$\mathcal{C} \triangleq \bigcup_{i,j\in[n],t\in[m]} \mathcal{C}_{ijt} = \left\{\mathbf{v} \in \mathbb{R}^{rd_2} : \prod_{i,j\in[n],t\in[m]} (F_{i,j}(\mathbf{v}) - \gamma(\mathbf{v})s_t) = 0\right\}.$$

For any two points $\mathbf{v}_1$ and $\mathbf{v}_2$ in a single connected component of the complement, $\mathcal{C}^{\mathsf{c}}$, $|\Phi_{ij}(\mathbf{v}_1) - \Phi_{ij}(\mathbf{v}_2)| \leq \delta$ for all $(i,j) \in [n] \times [n]$. Therefore, $\|\mathbf{P}_{\mathbf{v}_1,\mathcal{S}} - \mathbf{P}_{\mathbf{v}_2,\mathcal{S}}\|_{\mathrm{F}} \leq n\delta$. This implies that $N_{n\delta}(\mathcal{P}_{\mathcal{S}})$ is bounded by the number of connected components of $\mathcal{C}^{\mathsf{c}}$. Define

$$\Phi : \mathbb{R}^{rd_2+1} \to \mathbb{R}, (v_0, \mathbf{v}^\top)^\top \mapsto \left\{\gamma(\mathbf{v}) \times \prod_{i,j\in[n],t\in[m]} v_0(F_{i,j}(\mathbf{v}) - \gamma(\mathbf{v})s_t)\right\} - 1.$$

We have that $\Phi^{-1}(0)$ shares the same number of connected components as $\mathcal{C}^{\mathsf{c}}$. By Theorem 7.23 of Basu et al. (2007), the number of connected components of $\Phi^{-1}(0)$, which is the 0th Betti number of $\Phi^{-1}(0)$, is bounded by $\{(4|\mathcal{S}|-1)n^2m\}^{rd_2+1}$. Therefore, for any $\delta < 1$,

$$N_{n\delta}(\mathcal{P}) \leq \left\{(4|\mathcal{S}|-1)n^2\frac{3}{\delta}\right\}^{rd_2+1}.$$

Then we deduce that

$$N_{\delta}(\mathcal{P}_{\mathcal{S}}) \leq \left\{\frac{3(4|\mathcal{S}|-1)n^3}{\delta}\right\}^{rd_2+1}.$$

Finally, since $\mathcal{P} \subseteq \bigcup_{\mathcal{S}\subseteq[rd_1]} \mathcal{P}_{\mathcal{S}}$, we have

$$N_{\delta}(\mathcal{P}) \leq \sum_{\mathcal{S}\subseteq[rd_1]} N_{\delta}(\mathcal{P}_{\mathcal{S}}) \leq \sum_{\mathcal{S}\subseteq[rd_1]} \left\{\frac{3(4|\mathcal{S}|-1)n^3}{\delta}\right\}^{2rd_2+1} \leq 2^{rd_1}\left\{\frac{12rd_1n^3}{\delta}\right\}^{rd_2+1},$$

which concludes the proof. □

To bound the in-sample prediction risk with high-probability, we need the following mild assumption, which is standard in high-dimensional models. In order to state our assumption, we first define sub-Gaussian random variables.

Definition 5.2.1. For a random variable ξ valued in $\mathbb{R}$, define the ψ_2-norm of ξ, denoted $\|\xi\|_{\psi_2}$, as

$$\|\xi\|_{\psi_2} \triangleq \inf_{t>0}\{\mathbb{E}\exp(t^{-2}\xi^2) \leq 2\}.$$

Then, define the family of sub-Gaussian random variables with parameter K as

$$\mathcal{SG}(\varsigma) \triangleq \{\xi : \|\xi\|_{\psi_2} \leq K\}.$$

More generally, for p-dimensional real-valued random vectors, we define the sub-Gaussian family with parameter K as

$$\mathcal{SG}_p(\varsigma) \triangleq \Big\{\boldsymbol{\xi} : \sup_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \|\langle\boldsymbol{\xi}, \mathbf{v}\rangle_2\|_{\psi_2} \leq K\Big\}.$$

Assumption 5.1. The noise $\varepsilon \in \mathcal{SG}(K_\varepsilon)$ with mean zero and variance σ_ε^2 and is independent of $\mathbf{X}$.

Now, we can state our main result for $\hat{\mathbf{\Theta}}_{L_0}$.

Theorem 5.3. *Suppose we have n observations $(\mathbf{X}_i, y_i)_{i\in[n]}$ from model* (5.1.1) *with $(\varepsilon_i)_{i\in[n]}$ independent. Under Assumption* (5.1)*, if $r \geq r^*$, then there exist $c_1, c_2 > 0$ depending on σ_ε^2 and K_ε such that*

$$\mathcal{R}(\hat{\mathbf{\Theta}}_{L_0}) \leq c_1 \frac{rd\log(n)}{n}$$

with probability at least $1 - 4\exp(-c_2 rd\log(n))$.

Theorem 5.3 should be compared with the results of Rohde and Tsybakov (2011) and Koltchinskii et al. (2011), who proved bounds on in-sample prediction for the estimator $\hat{\mathbf{\Theta}}_{\mathrm{N}}$. Up to logarithmic factors, both $\hat{\mathbf{\Theta}}_{L_0}$ and $\hat{\mathbf{\Theta}}_{\mathrm{N}}$ achieve the same in-sample prediction risk; however, the crucial difference between our result and the existing results is the assumption, or lack thereof, on the design matrices, $\mathbf{X}_i$. The estimator $\hat{\mathbf{\Theta}}_{\mathrm{N}}$, much like the lasso estimator for linear models, requires a restricted eigenvalue type assumption in order to enjoy near optimal rates of in-sample prediction risk. By comparison, Theorem 5.3 imposes no such requirement.

In above theorem, we have assumed that the tuning parameter, r, exceeds the true rank, r^*. The following corollary extends the result to the risk bound to the setting where $r < r^*$.

Corollary 5.3.1. *Consider n observations from a signal-plus-noise model*

$$y = f + \varepsilon.$$

For $r > 0$, define

$$\hat{\mathbf{\Theta}}_{L_0}(r) \triangleq \operatorname*{arg\,min}_{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2}, \mathrm{rank}(\mathbf{\Theta})\leq r} \sum_{i=1}^n (y_i - \langle\mathbf{X}_i, \mathbf{\Theta}\rangle_{HS})^2.$$

Assume that ε_i are independent and identically distributed sub-Gaussian random variables with parameter K_ε. Then, there exist constants $c_1, c_2 > 0$ depending on σ_ε^2 and K_ε such that

$$\mathcal{R}(\hat{\mathbf{\Theta}}_{L_0}) \leq \Big\{\Big[\min_{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2}, \mathrm{rank}(\mathbf{\Theta})\leq r} \mathcal{R}(\mathbf{\Theta})\Big]^{1/2} + \Big[c_1\frac{rd\log(n)}{n}\Big]^{1/2}\Big\}^2$$

with probability at least $1 - 4\exp(-c_2 rd\log(n))$.

Compared with Theorem 5.3, Corollary 5.3.1 has an additional term denoting the best attainable risk using a rank r approximation; when $r \geq r^*$, the best approximation error is zero and we recover Theorem 5.3.

5.3 Testing in Signal Plus Noise Models

5.3.1 A General Power Analysis of a Permutation Test

Consider the following general signal-plus-noise model:

$$y = f(x) + \varepsilon \triangleq \mathbb{E}(y|x) + \varepsilon, \tag{5.3.1}$$

where x is a random covariate vector valued in $\mathscr{X}$. For simplicity, write $f(x) \triangleq \mathbb{E}(y|x)$. Then, we are interested in the hypotheses testing problem

$$H_0 : f \equiv 0 \qquad\qquad H_1 : f \in \mathscr{F}, f \not\equiv 0 \tag{5.3.2}$$

for some prespecified function class $\mathscr{F}$. We discuss some examples of $\mathscr{F}$ in Sections 5.3.2 – 5.3.3. To test these hypotheses, we consider a flexible permutation test. We note that the application of permutation tests to signal plus noise models is not novel. For example, the seminal work of Freedman and Lane (1983) proposed a permutation test for a collection of covariates in a low-dimensional linear model. However, to the best of our knowledge, the theory of permutation tests for high-dimensional models has not been explored, particularly the power of permutation tests. Under mild assumptions, such as exchangability of $(\varepsilon_i)_{i\in[n]}$ given $(x_i)_{i\in[n]}$, it is easy to derive a test statistic that controls type I error under the permutation null hypothesis. However, the sparsity rate only enters in the power under the alternative. In Theorem 5.6 below, we characterize explicitly this dependence of power on the sparsity.

Before presenting the test statistic, we need to establish some notation and facts from enumerative combinatorics regarding permutations that are used throughout the section. Let $\Pi = \Pi_n$ denote the set of all permutations over $[n]$. For a given $\pi \in \Pi$, we write $f^{(\pi)}(x_i) \triangleq \mathbb{E}(y_{\pi(i)}|x_i)$, noting that if $\pi(i) \neq i$, then $f^{(\pi)}(x_i) = 0$. We define π_0 to be the identity permutation on $[n]$. Moreover, a permutation $\pi \in \Pi$ induces a partition of $[n]$ into $K(\pi)$ cycles, where a cycle is $i_1, \ldots, i_k$ such that $\pi(i_j) = i_{j+1}$ for $j \in [k-1]$ and $\pi(i_k) = i_i$. Let

$$\tilde{\Pi} \triangleq \tilde{\Pi}_n = \{\pi \in \Pi : K(\pi) \leq \log^2(n)\}.$$

Suppose we have n independent and identically distributed observations $(x_i, y_i)_{i\in[n]}$ of (x, y) that follows model (5.3.1). We make the following assumptions.

Assumption 5.2. The mean $f(x)$ satisfies that $\mathbb{E}[f(x)] = 0$, $\mathbb{E}[f^2(x)] = \sigma_f^2$, and $\mathbb{E}[f^4(x)] < \infty$. The error ε satisfies that $\mathbb{E}(\varepsilon_1) = 0$ and $\mathbb{E}(\varepsilon_1^2) = \sigma_\varepsilon^2$ and is independent of $x_1, \ldots, x_n$.

Assumption 5.2*. The mean f satisfies that Var$(f^2(x)) \le \vartheta\sigma_f^{24}$ for some constant $\vartheta > 0$.

Assumption 5.3. For a fixed $\delta > 0$, there exists an estimator $\hat{f} : \mathscr{X} \times (\mathscr{X} \times \mathbb{R})^n \times \Pi \to \mathbb{R}$ and a sequence ℓ_n (possibly depending on δ) such that

(i) the estimator $\hat{f}$ is equivariant in the sense that for any $\pi \in \Pi$,

$$\hat{f}(x_i; (x_j, y_j)_{j=1}^n; \pi) = \hat{f}(x_i; (x_j, y_{\pi(j)})_{j=1}^n; \pi_0).$$

(ii) for n sufficiently large,

$$\min_{\pi\in\tilde{\Pi}\cup\{\pi_0\}} \mathbb{P}\Big\{ \sum_{i=1}^n [\hat{f}(x_i; (x_j, y_j)_{j=1}^n; \pi) - f^{(\pi)}(x_i)]^2 \le \ell_n \Big\} \ge 1 - \delta.$$

Temporarily fix $\pi \in \Pi$. For convenience, let $\hat{f}^{(\pi)}(\cdot) \triangleq \hat{f}(\cdot; (x_j, y_j)_{j=1}^n; \pi)$. Now, given an estimation procedure $\hat{f} : \mathscr{X} \times (\mathscr{X} \times \mathbb{R})^n \times \Pi \to \mathbb{R}$ satisfying Assumption (5.3), we define $\Lambda^{(\pi)}$ as

$$\Lambda^{(\pi)} \triangleq \Lambda^{(\pi)}(\hat{f}) = \sum_{i=1}^n [\hat{f}^{(\pi)}(x_i)]^2.$$

Then, our p-value is given by

$$\varphi(\hat{f}) \triangleq \frac{1}{|\Pi|} \sum_{\pi\in\Pi} \mathbb{1}_{\Lambda^{(\pi_0)}(\hat{f}) \le \Lambda^{(\pi)}(\hat{f})}.$$

Assumption (5.2) is a weak requirement, imposing an independence assumption and some moment conditions on the model. The requirement for the existence of the fourth moment of f_i is to ensure the concentration of $\|\boldsymbol{f}\|_2^2$ around $n\sigma_f^2$. The next assumption, (5.2*), is a technical condition that allows for a faster concentration of $\|\boldsymbol{f}\|_2^2$; the faster concentration yields a sharper rate in the contiguous alternative. For example, if the f_i are Gaussian, then Assumption (5.2*) is satisfied with $\vartheta = 3$.

Assumption (5.3) is a very natural assumption, albeit technical. For the first part, the symmetry in the estimation procedure, $\hat{f}(\cdot)$, implies that $\Lambda^{(\pi)}$ is identically distributed under the null hypothesis for all $\pi \in \Pi$. For the second half, we assume that $\hat{f}^{(\pi)}(\cdot)$ is a consistent estimator of $f^{(\pi)}(\cdot)$

for any $\pi \in \Pi$ at a rate slightly faster than ℓ_n. In particular, for most $\pi \in \Pi$, the estimator $\hat{f}^{(\pi)}(\cdot)$ approximates the zero function. To see this, let $C_1, \ldots, C_{K(\pi)}$ denote a fixed representation of the $K(\pi)$ cycles, for example as expressed in *standard representation* (see Stanley (2012) for a formal definition). For $j \in [K(\pi)]$ and $i \in C_j$, let $m(i)$ denote the index of i in C_j. Then, define the sets $\mathcal{A}_1^{(\pi)}$, $\mathcal{A}_2^{(\pi)}$, and $\mathcal{A}_3^{(\pi)}$ as follows:

$$\begin{aligned}
\mathcal{A}_1^{(\pi)} &\triangleq \bigcup_{j \in [K(\pi)]} \{i \in C_j : m(i) \text{ is odd and } m(i) \neq |C_j|\}, \\
\mathcal{A}_2^{(\pi)} &\triangleq \bigcup_{j \in [K(\pi)]} \{i \in C_j : m(i) \text{ is even}\}, \\
\mathcal{A}_3^{(\pi)} &\triangleq [n] \cap (\mathcal{A}_1^{(\pi)} \cup \mathcal{A}_2^{(\pi)})^{\mathsf{c}}.
\end{aligned}$$

For example, for the permutation expressed by the cycles $(4321)(765)(8)$, we have $\mathcal{A}_1^{(\pi)} = \{2, 4, 7\}$, $\mathcal{A}_2^{(\pi)} = \{1, 3, 6\}$, and $\mathcal{A}_3^{(\pi)} = \{5, 8\}$. Intuitively, $\mathcal{A}_1^{(\pi)}$ and $\mathcal{A}_2^{(\pi)}$ are two sets of observations such that, within each set, the covariates and responses are mutually independent. Therefore, for $i \in \mathcal{A}_1^{(\pi)} \cup \mathcal{A}_2^{(\pi)}$, we have that $f^{(\pi)}(x_i) = 0$. The other set, $\mathcal{A}_3^{(\pi)}$, are the remaining observations. To bound the error in the remaining observations, we note that $\mathbb{E}K(\pi)/\log(n) \to 1$ as $n \to \infty$ (cf. Stanley (2012)). Now, by Markov's inequality, for large values of n,

$$\frac{|\tilde{\Pi}^{\mathsf{c}}|}{|\Pi|} = \mathbb{P}(K(\pi) > \log^2(n)) \leq \frac{2}{\log(n)} \to 0.$$

This leads to the following lemma, which asserts that, for $\pi \in \tilde{\Pi}$, the conditional mean function, $f^{(\pi)}(\cdot)$, is approximately zero.

Lemma 5.4. *Under Assumption* (5.2)*, for any $\delta > 0$, there exists a constant $c_1 > 0$ such that*

$$\min_{\pi \in \tilde{\Pi}} \mathbb{P}\Big\{ \sum_{i \in \mathcal{A}_3^{(\pi)}} [f^{(\pi)}(x_i)]^2 \leq \log^2(n)\sigma_f^2 + c_1 \log(n) \Big\} \geq 1 - \delta.$$

As an immediate consequence of Lemma 5.4, we have the following corollary, which yields an alternative way to check the second half of Assumption (5.3).

Suppose that, for a fixed $\delta > 0$, there exists an estimator $\hat{f} : \mathscr{X} \times (\mathscr{X} \times \mathbb{R})^n \times \Pi \to \mathbb{R}$ and a sequence ℓ_n with $\log^2(n) = o(\ell_n)$ such that for n sufficiently large,

$$\mathbb{P}\Big\{ \sum_{i=1}^{n} [\hat{f}(x_i; (x_j, y_j)_{j=1}^n; \pi_0) - f^{(\pi_0)}(x_i)]^2 \leq \ell_n \Big\} \geq 1 - \delta$$

and

$$\min_{\pi \in \tilde{\Pi}} \mathbb{P}\Big\{ \sum_{i=1}^{n} [\hat{f}(x_i; (x_j, y_j)_{j=1}^n; \pi)]^2 \leq \ell_n \Big\} \geq 1 - \delta.$$

Then, under Assumption (5.2), for n sufficiently large,

$$\min_{\pi \in \tilde{\Pi} \cup \{\pi_0\}} \mathbb{P}\Big\{ \sum_{i=1}^{n} [\hat{f}(x_i; (x_j, y_j)_{j=1}^n; \pi) - f^{(\pi)}(x_i)]^2 \leq \ell_n \Big\} \geq 1 - \delta.$$

We can now state our first result that φ controls the type-I error.

Theorem 5.5. *Consider model* (5.3.1) *with the hypotheses testing problem in* (5.3.2)*. Under Assumptions* (5.2)*,* (5.3) *and the null hypothesis, we have that*

$$\limsup_{n \to \infty} \mathbb{P}_{H_0}(\varphi(\hat{f}) \leq \alpha) \leq \alpha.$$

To analyze the power of the test, we consider two contiguous hypotheses testing problems, depending on whether we impose Assumption (5.2*). First, consider

$$H_0 : f \equiv 0 \qquad \text{v.} \qquad H_1 : f \not\equiv 0, f \in \mathscr{F}, \sigma_f^2 = h\Big(\frac{1}{\sqrt{n}} + \frac{\ell_n}{n}\Big). \tag{5.3.3}$$

Then, we have the following theorem that

Theorem 5.6. *Consider model* (5.3.1) *with the hypotheses testing problem* (5.3.3)*. Suppose Assumptions* (5.2) *and* (5.3) *hold with $\delta < \alpha(1-\alpha)/4$. If h is sufficiently large (not depending on n), then, under the alternative hypothesis in* (5.3.3)*,*

$$\liminf_{n \to \infty} \mathbb{P}_{H_1}(\varphi(\hat{f}) \leq \alpha) > \alpha.$$

If we further assume (5.2*), we consider the following hypotheses

$$H_0 : f \equiv 0 \qquad \text{v.} \qquad H_1 : f \not\equiv 0, f \in \mathscr{F}, \sigma_f^2 = h\frac{\ell_n}{n}. \tag{5.3.4}$$

We have the following corollary.

Corollary 5.6.1. *Consider model* (5.3.1) *with the hypotheses testing problem in* (5.3.4)*. Under Assumptions* (5.2)*,* (5.2*)*,* (5.3) *with $\delta < \alpha(1-\alpha)/4$ and the alternative hypothesis in* (5.3.4)*, if h*

is sufficiently large (not depending on n), then we have that

$$\liminf_{n\to\infty} \mathbb{P}_{H_1}(\varphi(\hat{f}) \leq \alpha) > \alpha.$$

Comparing the hypotheses in equations (5.3.3) and (5.3.4), one can see that Assumption (5.2*) allows for testing at a rate faster than $n^{-1/2}$. In view of Corollary 5.3.1, we emphasize that the bottleneck of testing power under Assumption (5.2*) is the rate at which we can predict the conditional mean given the permuted covariates.

5.3.2 Sparse High-Dimensional Linear Model

In this section, we focus on model (A.3.1.1) and consider $\mathscr{F} \triangleq \{f : \mathbb{R}^p \to \mathbb{R} \mid f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta}^* \rangle_2, \boldsymbol{\beta}^* \in \mathbb{R}^p, \|\boldsymbol{\beta}^*\|_0 \leq s^*\}$.

Assumption 5.4. The covariate vector $\mathbf{x} \in \mathcal{SG}_p(K_x)$ has mean zero and variance $\boldsymbol{\Sigma}$, and the error $\varepsilon \in \mathcal{SG}(K_\varepsilon)$ has mean zero and variance σ_ε^2. Moreover, $\mathbf{x}$ is independent of ε.

Assumption 5.4*. There exists $\vartheta > 0$ such that

$$\|\langle \mathbf{x}, \mathbf{v} \rangle_2\|_{\psi_2}^2 \leq \vartheta \mathbb{E}\big(\langle \mathbf{x}, \mathbf{v} \rangle_2^2\big)$$

for all $\mathbf{v} \in \mathbb{R}^p$.

The first half of Assumption (5.4) is mild, assuming a random sub-Gaussian design framework that is standard in the high-dimensional setting. In particular, it implies Assumption (5.2). Assumption (5.4*) is a technical assumption that is used in the literature for concentration of the sample covariance matrix. For example, see Definition 2 of Koltchinskii and Lounici (2017) or Theorem 4.7.1 of Vershynin (2018). In particular, it implies Assumption (5.2*), and, as an example, the Gaussian distribution satisfies this assumption.

Then, the pairs of contiguous testing problems that we consider are

$$H_0 : \boldsymbol{\beta}^* = \mathbf{0}_p \qquad \text{v.} \qquad H_1 : \|\boldsymbol{\beta}^*\|_0 = s^* > 0, (\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma} \boldsymbol{\beta}^* = h\Big(\frac{1}{\sqrt{n}} + \frac{s^* \log(p)}{n}\Big) \tag{5.3.5}$$

and, under Assumption (5.4*),

$$H_0 : \boldsymbol{\beta}^* = \mathbf{0}_p \qquad \text{v.} \qquad H_1 : \|\boldsymbol{\beta}^*\|_0 = s^* > 0, (\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma} \boldsymbol{\beta}^* = h\frac{s^* \log(p)}{n}. \tag{5.3.6}$$

Now, for any $\pi \in \Pi$, we define the lasso estimator as

$$\hat{f}_{\mathrm{LA}}(\mathbf{x}_i; (\mathbf{x}_j, y_j)_{j=1}^n; \pi) \triangleq \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{LA}}^{(\pi)} \rangle_2,$$

where

$$\hat{\boldsymbol{\beta}}_{\mathrm{LA}}^{(\pi)} \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{n}\sum_{i=1}^{n}(y_{\pi(i)} - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle_2)^2 + \lambda\|\boldsymbol{\beta}\|_1\Big\}. \tag{5.3.7}$$

Then, we have the following result for the lasso estimator.

Theorem 5.7. *Consider model* (A.3.1.1). *Suppose that Assumption* (5.4) *holds with* $0 < \lambda_{min}(\boldsymbol{\Sigma}) \leq \lambda_{max}(\boldsymbol{\Sigma}) < \infty$. *Then for a fixed value of* $\delta > 0$, *there exists sequence* $\ell_n = O(sn\lambda^2)$ *such that the lasso estimator* $\hat{f}_{LA}$ *satisfies (i) and (ii) of Assumption* (5.3), *provided that the tuning parameter* λ *in* (5.3.7) *satisfies* $\lambda \geq 2\lambda_0$ *and* $\lambda \asymp \lambda_0$, *where*

$$\lambda_0 \geq c_1\sqrt{K_x(K_f + K_\varepsilon)\frac{\log(6/\delta) + \log(p)}{n}}$$

for some universal constant $c_1 > 0$.

Given Theorem 5.7, applying Theorems 5.5, 5.6 and Corollary 5.6.1 yields the following corollary on the asymptotic validity of the permutation test based on $\hat{f}_{\mathrm{LA}}$.

Corollary 5.7.1. *Under the assumptions of Theorem 5.7,*

$$\limsup_{n\to\infty} \mathbb{P}_{H_0}(\varphi(\hat{f}_{LA}) \leq \alpha) \leq \alpha$$

In addition, if h *is sufficiently large (not depending on* n*), then*

$$\liminf_{n\to\infty} \mathbb{P}_{H_1}(\varphi(\hat{f}_{LA}) \leq \alpha) > \alpha$$

for the hypotheses testing problem in equation (5.3.5) *and also for the hypotheses in equation* (5.3.6) *if Assumption* (5.4*) *holds.*

Similarly, we define the L_0 estimator as

$$\hat{f}_{L_0}(\mathbf{x}_i; (\mathbf{x}_j, y_j)_{j=1}^n; \pi) \triangleq \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\pi)}\rangle_2,$$

where

$$\hat{\boldsymbol{\beta}}_{L_0}^{(\pi)} \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \|\boldsymbol{\beta}\|\leq s}{\arg\min}\sum_{i=1}^{n}(y_{\pi(i)} - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle_2)^2. \tag{5.3.8}$$

Now, the following theorem is the analogue of Theorem 5.7 for the L_0 estimator.

Theorem 5.8. *Consider model* (A.3.1.1). *Suppose that* $s \asymp s^*$ *with* $s \geq s^*$ *in* (5.3.8). *Then under Assumption* (5.4), *for a fixed value of* $\delta > 0$, *the* L_0 *estimator,* $\hat{f}_{L_0}(\cdot)$, *satisfies Assumption* (5.3) *with* $\ell_n = O(s\log(p) + \log(1/\delta))$.

It is worth emphasis that Theorem 5.8 does not require the minimum eigenvalue of $\mathbf{\Sigma}$ to be well bounded from below as in Theroem 5.7. This demonstrates the robustness of the L_0 estimator against collinearity of the covariates when it is compared with the lasso. In the following, we establish the asymptotic validity of the permutation test based on $\hat{f}_{L_0}$, again without any requirement on $\lambda_{\min}(\mathbf{\Sigma})$.

Corollary 5.8.1. *Under the assumptions of Theorem 5.8, then*

$$\limsup_{n\to\infty} \mathbb{P}_{H_0}(\varphi(\hat{f}_{L_0}) \leq \alpha) \leq \alpha$$

In addition, if h is sufficiently large (not depending on n), then

$$\liminf_{n\to\infty} \mathbb{P}_{H_1}(\varphi(\hat{f}_{L_0}) \leq \alpha) > \alpha$$

for the hypotheses testing problem in equation (5.3.5) *and also for the hypotheses in equation* (5.3.6) *if Assumption* (5.4*) *holds.*

Remark. If $\mathbf{\Sigma} = \mathbf{I}_p$, then we are interested in testing if $\|\boldsymbol{\beta}^*\|_2^2 = 0$. Our results should be compared to the minimax lower bound of Guo et al. (2019), who show that the minimax lower bound of the estimation error of $\|\boldsymbol{\beta}^*\|_2^2$ is $n^{-1/2} + s^* n^{-1}\log(p)$ over all s^*-sparse vectors with bounded Euclidean norms. However, under Assumption (5.4*), we are able to test at a faster rate since $\boldsymbol{\beta}^* = \mathbf{0}_p$ is a super-efficient point in the parameter space.

5.3.3 Low-Rank Trace Regression

Now we return to the main subject of this paper, the low-rank trace regression model (5.1.1). Here, we let $\mathscr{F} \triangleq \{f : \mathbb{R}^{d_1\times d_2} \to \mathbb{R} \mid f(\mathbf{X}) = \langle \mathbf{X}, \mathbf{\Theta}\rangle_{\text{HS}}, \operatorname{rank}(\mathbf{\Theta}) \leq r^*\}$. Similarly to the setting of high-dimensional linear models, we require the following mild assumption on the covariates and noise.

Assumption 5.5. The vectorized covariate matrix $\operatorname{vec}(\mathbf{X}) \in \mathcal{SG}_{d_1 d_2}(K_x)$ with mean zero and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$. The error $\varepsilon \in \mathcal{SG}(K_\varepsilon)$ and has mean zero and variance σ_ε^2. Moreover, $\mathbf{X}$ is independent of ε.

Assumption 5.5*. There exists a $\vartheta > 0$ such that

$$\|\langle \operatorname{vec}(\mathbf{X}), \mathbf{v}\rangle_2\|_{\psi_2}^2 \leq \vartheta \mathbb{E}\langle \operatorname{vec}(\mathbf{X}), \mathbf{v}\rangle_2^2$$

for all $\mathbf{v} \in \mathbb{R}^{d_1 d_2}$.

The corresponding two pairs of contiguous hypotheses we consider for the low-rank trace regression model are

$$H_0 : \mathbf{\Theta}^* = \mathbf{0}_{d_1 \times d_2} \qquad \text{v.} \qquad H_1 : \operatorname{rank}(\mathbf{\Theta}^*) = r^* > 0, \\ \operatorname{vec}(\mathbf{\Theta}^*)^\mathsf{T} \mathbf{\Sigma} \operatorname{vec}(\mathbf{\Theta}^*) = h\Big(\frac{1}{\sqrt{n}} + \frac{r^* d \log(n)}{n}\Big) \tag{5.3.9}$$

and

$$H_0 : \mathbf{\Theta}^* = \mathbf{0}_{d_1 \times d_2}, \qquad \text{v.} \qquad H_1 : \operatorname{rank}(\mathbf{\Theta}^*) = r^* > 0, \operatorname{vec}(\mathbf{\Theta}^*)^\mathsf{T} \mathbf{\Sigma} \operatorname{vec}(\mathbf{\Theta}^*) = h\frac{r^* d \log(n)}{n}. \tag{5.3.10}$$

For any $\pi \in \Pi$, we define the rank-constrained estimator as

$$\hat{f}_{L_0}(\mathbf{X}_i; (\mathbf{X}_j, y_j)_{j=1}^n; \pi) \triangleq \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}^{(\pi)} \rangle_{\text{HS}},$$

where

$$\hat{\mathbf{\Theta}}_{L_0}^{(\pi)} \triangleq \hat{\mathbf{\Theta}}_{L_0}^{(\pi)}(r) = \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \operatorname{rank}(\mathbf{\Theta}) \le r}{\arg\min} \sum_{i=1}^n (y_{\pi(i)} - \langle \mathbf{X}_i, \mathbf{\Theta} \rangle_{\text{HS}})^2. \tag{5.3.11}$$

Next, we show that $\hat{f}_{L_0}(\cdot)$ satisfies Assumption C without any requirement on $\mathbf{\Sigma}$.

Theorem 5.9. *Suppose that Assumption* (5.5) *holds and that* $r \asymp r^*$ *with* $r \ge r^*$ *in* (5.3.11)*. Then for some* $\delta > 0$*,* $\hat{f}_{L_0}$ *satisfies (i) and (ii) of Assumption* (5.3) *with some* $\ell_n = O(r^* \log(d_1 d_2) + \log(1/\delta))$.

We can now establish the asymptotic validity of the low-rank test, again through applying Theorems 5.5, 5.6 and Corollary 5.6.1.

Corollary 5.9.1. *Under the assumptions of Theorem 5.9, we have that*

$$\limsup_{n \to \infty} \mathbb{P}_{H_0}(\varphi(\hat{f}_{L_0}) \le \alpha) \le \alpha.$$

In addition, if h *is sufficiently large (not depending on* n*), then*

$$\liminf_{n \to \infty} \mathbb{P}_{H_1}(\varphi(\hat{f}_{L_0}) \le \alpha) > \alpha$$

for the hypotheses testing problem in equation (5.3.9) *and also for the hypotheses in equation* (5.3.10) *if Assumption* (5.5*) *holds.*

5.3.4 Robustness of the Rank-Constrained Test

In the previous sections, we assume that our test statistics have been optimally tuned, either through λ for regularized estimation or through r for rank-constrained estimation. However, such oracles are not available in practice. In this section, we consider the performance of the permutation test with a possibly misspecified rank. Fix r and define

$$\tilde{\boldsymbol{\Theta}} \triangleq \tilde{\boldsymbol{\Theta}}(r) = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(\boldsymbol{\Theta}) \leq r}{\arg\min} \mathbb{E}\langle \mathbf{X}, \boldsymbol{\Theta}^* - \boldsymbol{\Theta} \rangle_{\mathrm{HS}}^2.$$

In words, $\tilde{\boldsymbol{\Theta}}$ is the best rank-r approximation to $\boldsymbol{\Theta}^*$ in terms of prediction risk if $1 \leq r < r^*$ and $\tilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^*$ if $r \geq r^*$. Then given $h > 0$, consider the hypotheses testing problems

$$H_0 : \boldsymbol{\Theta}^* = \mathbf{0}_{d_1 \times d_2} \qquad \text{v.} \qquad H_1 : \mathrm{vec}(\tilde{\boldsymbol{\Theta}})^\mathsf{T} \boldsymbol{\Sigma} \mathrm{vec}(\tilde{\boldsymbol{\Theta}}) = h\Big(\frac{1}{\sqrt{n}} + \frac{rd\log(n)}{n}\Big) \tag{5.3.12}$$

and

$$H_0 : \boldsymbol{\Theta}^* = \mathbf{0}_{d_1 \times d_2} \qquad \text{v.} \qquad H_1 : \mathrm{vec}(\tilde{\boldsymbol{\Theta}})^\mathsf{T} \boldsymbol{\Sigma} \mathrm{vec}(\tilde{\boldsymbol{\Theta}}) = h\frac{rd\log(d)}{n}. \tag{5.3.13}$$

Note that the alternative hypotheses above are stated in terms of $\tilde{\boldsymbol{\Theta}}$ and thus vary with respect to $\tilde{r}$. Intuitively, underestimating the rank refrains one from capturing the complete signal. It is thus hopeless to detect the presence of a nonzero $\boldsymbol{\Theta}^*$ if the signal encapsulated by $\tilde{\boldsymbol{\Theta}}$ is too weak.

Theorem 5.10. *Consider model* (5.1.1) *and choose $r = r$ in* (5.3.11). *Under Assumption* (5.5)*, we have that*

$$\limsup_{n\to\infty} \mathbb{P}_{H_0}(\varphi(\hat{f}_{L_0}) \leq \alpha) \leq \alpha.$$

In addition, if h is sufficiently large (not depending on n), then

$$\liminf_{n\to\infty} \mathbb{P}_{H_1}(\varphi(\hat{f}_{L_0}) \leq \alpha) > \alpha$$

for the hypotheses testing problem in equation (5.3.12) *and also for the hypotheses in equation* (5.3.13) *if Assumption* (5.5*) *holds.*

Theorem 5.10 should be compared with Corollary 5.9.1. In particular, by considering $\tilde{\boldsymbol{\Theta}}$ rather than $\boldsymbol{\Theta}^*$, we can still distinguish between the null and the alternative hypotheses as long as the best rank-r approximation captures sufficient amount of signal; hence, this allows for the situation where the rank of $\boldsymbol{\Theta}^*$ is misspecified. In particular, by setting $r = r^*$, the hypotheses in equations (5.3.12) and (5.3.13) are equivalent to equations (5.3.9) and (5.3.10) respectively, and Corollary

5.9.1 can be viewed as a special case of Theorem 5.10. As another special case, by setting $r = 1$, we obtain a tuning-parameter free test that allows for testing the best rank-one approximation of $\mathbf{\Theta}^*$. To the best of our knowledge, this is the first test in the high-dimensional literature that is robust to misspecification of the tuning parameter.

Theorem 5.10 seems to imply that the test is more likely to detect the signal if $\tilde{r}$ is large. However, it should be noted that the required minimal power depends linearly on r while the signal increases at most by a factor of r. Thus, the rank-one test, being focused on the leading eigenvalue, may have higher efficiency than a test that is more omnidirectional (for example, see Bickel et al. (2006)).

The proof of Theorem 5.10 relies on the least-squares structure of the rank-constrained estimator; in particular, the vector of fitted values can be written as $\mathbf{P}_{\mathbf{V}}\mathbf{y}$ for some $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$. Thus, the result can immediately be extended to sparse high-dimensional linear models with best subset selection.

By comparison, the choice of the tuning parameter λ for the lasso and nuclear norm regularized estimator is inherently challenging. For estimation, the value of λ is usually chosen through cross-validation. However, for inference, there are a few natural methods to perform cross-validation, which we discuss in Section 5.4.3.

5.4 Simulations

5.4.1 Models and Methods

In this section, we demonstrate the empirical performance, both in terms of estimation and inference, of the rank-constrained estimator on synthetic data. We assume the model in equation (5.1.1), which is reproduced below

$$y_i = \langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle_{\text{HS}} + \varepsilon_i.$$

In our simulations, we set $n = 200$ and $d_1 = d_2 = 20$ and let $r^* \in \{1, 2, 3, 4\}$. Regarding the design, we consider two distinct settings, corresponding to two common examples of low-rank trace regression: (i) compressed sensing and (ii) matrix completion. In the setting of compressed sensing, we let $\text{vec}(\mathbf{X}_i)$ have independent and identically distributed standard Gaussian entries. For matrix completion, we let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be a uniform random sample from $\{\mathbf{e}_j \mathbf{e}_k^\mathsf{T}\}_{j \in [d_1], k \in [d_2]}$ without replacement, where $\mathbf{e}_j$ denotes the jth standard basis vector.

In both scenarios, we generate ε_i as independent and identically distributed standard Gaussian random variables. Finally, we define the signal to noise ratio, denoted by "SNR," as the variance of $\langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle_{\text{HS}}$; the value of SNR is a monotonic function of the power represented by h in equations (5.3.3) and (5.3.4). For in-sample prediction, we consider a logarithmic scale and let $\text{SNR} \in \{1, 1.43, 2.04, 2.92, 4.18, 5.98, 8.55, 12.23, 17.48, 25\}$ and, for inference, we let $\text{SNR} \in \{0, 0.125, 0.25, 0.375, 0.5, 0.75, 1, 2\}$. To achieve this, we first generate r^* values uniformly from $(-1, 1)$ to form a diagonal matrix $\mathbf{\Lambda}$. Then, we draw $\mathbf{U}^*$ and $\mathbf{V}^*$ from the uniform Haar measure on the Stiefel manifold of dimension $d_1 \times r^*$ and $d_2 \times r^*$ respectively and set $\mathbf{\Theta}^* = \mathbf{U}^* \mathbf{\Lambda} (\mathbf{V}^*)^{\mathsf{T}}$. Finally, we scale $\mathbf{\Theta}^*$ such that $\text{vec}(\mathbf{\Theta}^*)^{\mathsf{T}} \mathbf{\Sigma} \text{vec}(\mathbf{\Theta}^*) = \text{SNR}$.

5.4.2 In-Sample Prediction

For estimation, we compare the in-sample prediction risk of the rank-constrained estimator with that of an oracle least-squares estimator (LS) and the nuclear norm regularized estimator (NN) from equation (5.1.4). The oracle least-squares estimator has access to the right singular space $\mathbf{V}^*$. Computationally, we use alternating minimization (AM) to approximate the rank-constrained estimator (for example, see Hastie et al. (2015) and the references therein). We employ multiple restarts to avoid local stationary points, using a coarse grid of nuclear norm estimators and a spectral estimator to initialize the AM algorithm; this yields a total of six initializations. Then, our final rank-constrained estimator is the one that minimizes equation (5.1.3) out of the six different initializations. To avoid misspecification of the tuning parameter for both estimators (r for the rank-constrained estimator and λ for the nuclear norm estimator), we consider oracle tuning parameters. To accomplish this, we run both estimators over a grid of tuning parameters for each setting over 1000 Monte Carlo experiments and choose the tuning parameter that yields the minimum prediction risk, which is defined as in (5.2.1).

The results of our simulation are presented in Tables 5.1 and 5.2 and Figures 5.2 and 5.3. In general, we see that the performance of the rank-constrained estimator relative to the nuclear norm regularized estimator improves as SNR increases. This is consistent with the simulation results of Hastie et al. (2020), who noticed that best subset selection outperforms the lasso for high SNR regimes in high-dimensional linear models.

5.4.3 Inference

For inference, we evaluate the performance of our permutation approach from Section 5.3 and consider the permutation test using both alternating minimization and nuclear norm regularization. Throughout, we are testing at level $\alpha = 0.05$, and our permutation tests randomly draw nineteen permutations from $\Pi \backslash \{\pi_0\}$. To provide a benchmark for the performance of our testing procedure, we consider two oracles that have access to the right singular space $\mathbf{V}^*$: (i) an oracle that uses the low-dimensional F-test (FT) and (ii) an oracle that uses our permutation test using least-squares

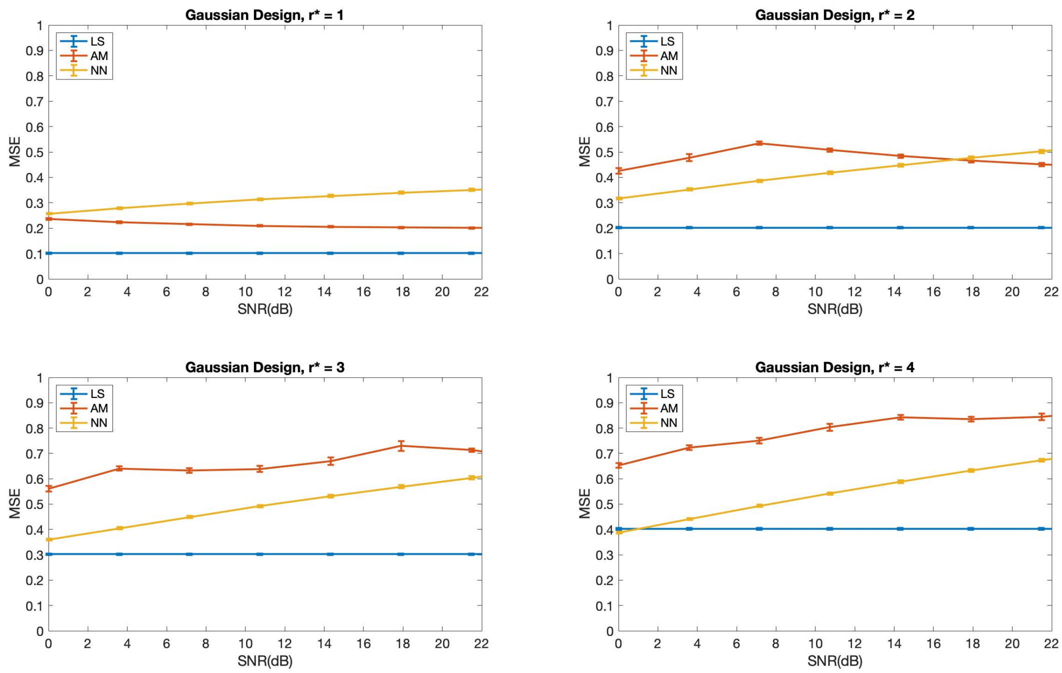

Figure 5.2: Plots of in-sample prediction error for Gaussian design

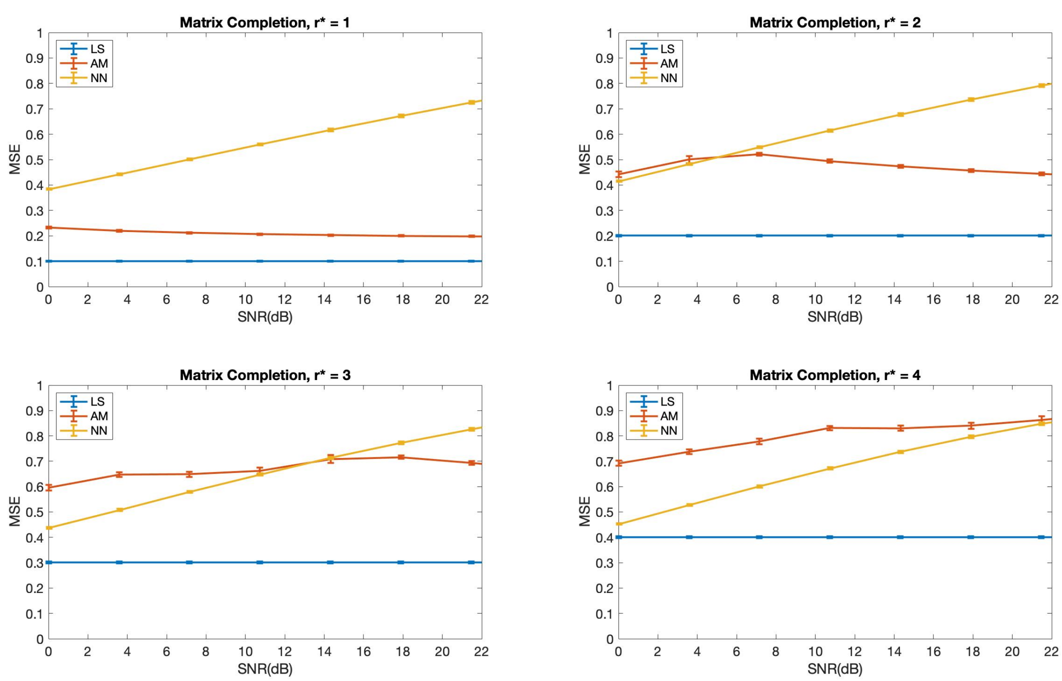

Figure 5.3: Plots of in-sample prediction error for matrix completion

Table 5.1: Simulations for In-Sample Prediction Risk for Gaussian Design

	SNR	1.00	1.43	2.04	2.92	4.18	5.98	8.55	12.23	17.48	25.00
	LS	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$r^* = 1$	AM	0.24	0.22	0.22	0.21	0.21	0.20	0.20	0.20	0.20	0.20
	NN	0.26	0.28	0.30	0.31	0.33	0.34	0.35	0.36	0.37	0.38
	LS	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
$r^* = 2$	AM	0.43	0.48	0.53	0.51	0.48	0.47	0.45	0.44	0.43	0.42
	NN	0.32	0.35	0.39	0.42	0.45	0.48	0.50	0.53	0.55	0.57
	LS	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
$r^* = 3$	AM	0.56	0.64	0.63	0.64	0.67	0.73	0.71	0.69	0.67	0.65
	NN	0.36	0.41	0.45	0.49	0.53	0.57	0.60	0.64	0.67	0.69
	LS	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
$r^* = 4$	AM	0.65	0.72	0.75	0.80	0.84	0.84	0.84	0.87	0.90	0.88
	NN	0.39	0.44	0.49	0.54	0.59	0.63	0.67	0.71	0.75	0.78

estimation (LS).

For nuclear norm regularization, we consider four procedures to choose λ. First, we consider oracle tuning. Since we use 100 Monte Carlo experiments, when choosing the oracle value of λ for nuclear norm regularization, we only consider the values of λ for which the number of rejections under the null hypothesis ($\mathrm{SNR} = 0$) is less than or equal to nine. The nine arises from constructing a confidence interval for α based on 100 independent Bernoulli experiments with success probability 0.05 as it is two standard errors above 0.05. The other three approaches to choosing λ are all variations on cross-validation. The first procedure, denoted DS for "data splitting," splits the data into two halves, estimating λ on the first half and using estimated value of λ for all $\pi \in \Pi$. In our simulations, we split the data in half since we need sufficient observations in both halves to estimate rank three and rank four matrices. The second procedure, denoted IS for "in-sample," performs five-fold cross-validation for each $\pi \in \Pi$ to obtain $\hat{\lambda}^{(\pi)}$. After $\hat{\lambda}^{(\pi)}$ is selected, the model is refit using all the observations to obtain in-sample predictions. Finally, the third procedure, denoted OS for "out-of-sample," also performs five-fold cross-validation for each $\pi \in \Pi$, obtaining five values of $\hat{\lambda}^{(\pi)}$, one for each of the five folds. Instead of refitting the model as before, we use the out-of-sample predicted values for each of the folds.

For alternating minimization, we report all the results for $r \in \{1, 2, 3, 4\}$. Note that we are using the oracle value of λ for nuclear norm regularization. Thus, we view this as a theoretical benchmark with which to compare the rank-constrained estimator for testing.

The results are presented in Tables 5.3 and 5.4 and Figures 5.4 and 5.5. We note that, as the SNR increases for a fixed rank, the power of our testing procedure increases. In general, even under misspecification of the tuning parameter for the rank-constrained estimator, we are able to maintain nominal coverage. Moreover, even when $r^* > 1$, it seems that the rank-constrained estimator with

Table 5.2: Simulations for In-Sample Prediction Risk for Matrix Completion

	SNR	1.00	1.43	2.04	2.92	4.18	5.98	8.55	12.23	17.48	25.00
	LS	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$r^* = 1$	AM	0.23	0.22	0.21	0.21	0.20	0.20	0.20	0.20	0.20	0.20
	NN	0.38	0.44	0.50	0.56	0.62	0.67	0.73	0.77	0.82	0.86
	LS	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
$r^* = 2$	AM	0.44	0.50	0.52	0.49	0.47	0.46	0.44	0.43	0.43	0.42
	NN	0.42	0.48	0.55	0.61	0.68	0.74	0.79	0.84	0.88	0.92
	LS	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
$r^* = 3$	AM	0.60	0.65	0.65	0.66	0.71	0.72	0.69	0.67	0.65	0.64
	NN	0.44	0.51	0.58	0.65	0.71	0.77	0.83	0.87	0.91	0.94
	LS	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
$r^* = 4$	AM	0.69	0.74	0.78	0.83	0.83	0.84	0.86	0.89	0.87	0.86
	NN	0.45	0.53	0.60	0.67	0.74	0.80	0.85	0.89	0.93	0.95

$r = 1$ has comparable performance to the optimal nuclear norm regularized estimator as well as permutation testing with larger values of r. Thus, even without any oracular knowledge of $\mathbf{\Theta}^*$, we may obtain a valid and powerful test that is tuning parameter free by using the rank-constrained estimator with $r = 1$, which is consistent with Theorem 5.10.

However, when λ is chosen via cross-validation, the performance of the nuclear-norm regularized estimator degrades significantly relative to the oracle. For data-splitting, which has the best empirical performance for non-oracle nuclear norm regularization, we lose half of our observations to selecting λ and, compared to the cross-fitting of Chernozhukov et al. (2018b), we cannot switch the roles of the two halves of the dataset. For the remaining two settings, where $\hat{\lambda}$ depends on $\pi \in \Pi$, we notice that $\hat{\lambda}^{(\pi)} < \hat{\lambda}^{(\pi_0)}$ for $\pi \neq \pi_0$. Thus, $\Lambda^{(\pi)}$ compensates the poorer model fit compared to $\Lambda^{(\pi_0)}$ by increasing the complexity of the model, thus enabling overfitting.

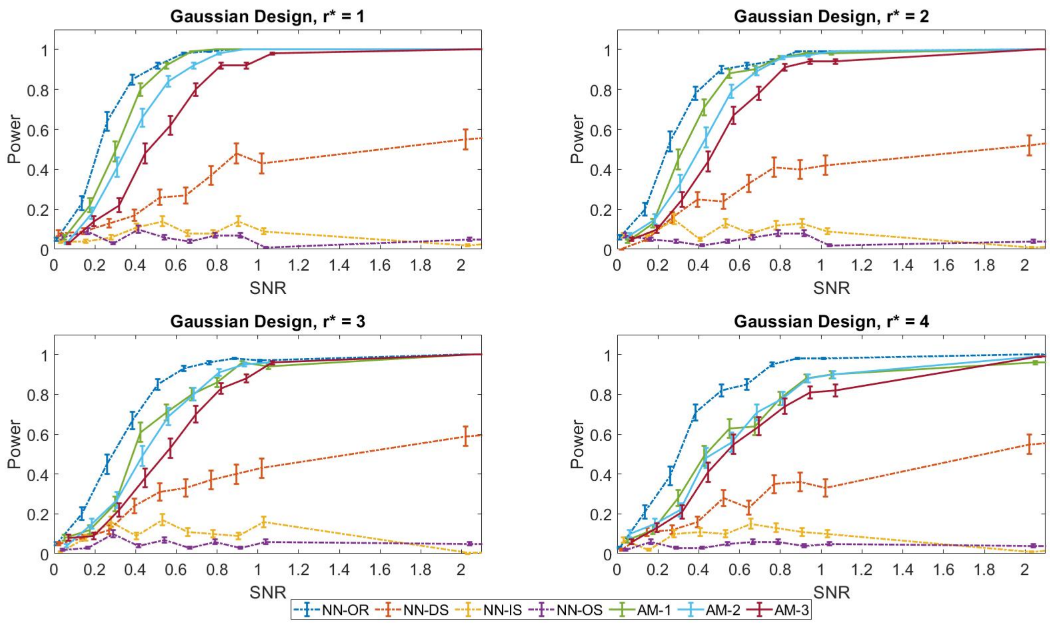

Figure 5.4: Plots of power for Gaussian design

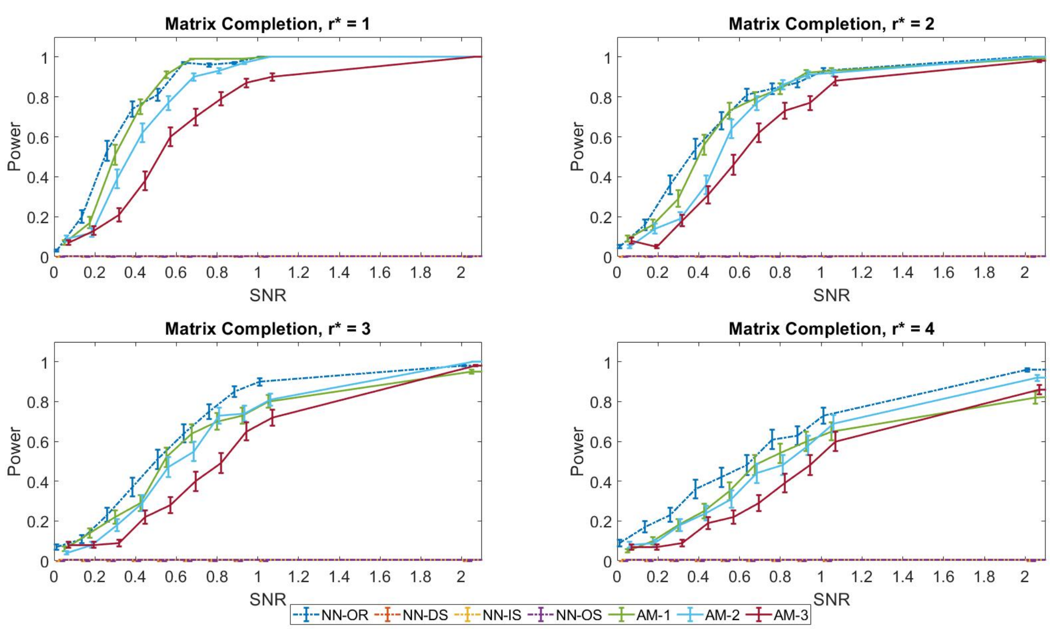

Figure 5.5: Plots of power for matrix completion

Table 5.3: Simulations for Inference for Gaussian Design

	SNR	0.000	0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000	2.000
	FT	0.02	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.06	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.05	0.23	0.64	0.85	0.92	0.98	0.99	1.00	1.00	1.00
	NN-DS	0.08	0.09	0.13	0.17	0.26	0.27	0.37	0.48	0.43	0.55
	NN-IS	0.04	0.04	0.06	0.11	0.14	0.08	0.08	0.14	0.09	0.02
$r^* = 1$	NN-OS	0.06	0.09	0.03	0.10	0.06	0.04	0.07	0.07	0.01	0.05
	AM-1	0.06	0.22	0.49	0.80	0.92	0.99	1.00	1.00	1.00	1.00
	AM-2	0.03	0.18	0.41	0.66	0.84	0.92	0.98	1.00	1.00	1.00
	AM-3	0.03	0.14	0.22	0.48	0.62	0.80	0.92	0.92	0.98	1.00
	AM-4	0.07	0.10	0.15	0.28	0.45	0.60	0.72	0.75	0.83	1.00
	FT	0.02	0.65	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.01	0.54	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.06	0.20	0.54	0.78	0.90	0.92	0.94	0.99	0.99	1.00
	NN-DS	0.00	0.05	0.16	0.25	0.24	0.33	0.41	0.40	0.42	0.52
	NN-IS	0.07	0.07	0.15	0.05	0.13	0.08	0.12	0.13	0.09	0.01
$r^* = 2$	NN-OS	0.07	0.05	0.04	0.02	0.04	0.06	0.08	0.08	0.02	0.04
	AM-1	0.04	0.13	0.45	0.71	0.88	0.90	0.96	0.98	0.98	1.00
	AM-2	0.07	0.15	0.33	0.56	0.79	0.89	0.96	0.97	0.99	1.00
	AM-3	0.05	0.10	0.25	0.44	0.67	0.78	0.91	0.94	0.94	1.00
	AM-4	0.07	0.10	0.18	0.25	0.41	0.54	0.64	0.78	0.77	1.00
	FT	0.02	0.56	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.03	0.48	0.84	0.96	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.05	0.20	0.45	0.67	0.85	0.93	0.96	0.98	0.97	1.00
	NN-DS	0.05	0.08	0.12	0.24	0.31	0.33	0.37	0.40	0.43	0.59
	NN-IS	0.01	0.08	0.16	0.09	0.17	0.11	0.10	0.09	0.16	0.00
$r^* = 3$	NN-OS	0.02	0.03	0.10	0.04	0.07	0.03	0.06	0.03	0.06	0.05
	AM-1	0.08	0.12	0.25	0.61	0.71	0.80	0.86	0.96	0.94	1.00
	AM-2	0.04	0.15	0.27	0.49	0.69	0.80	0.91	0.95	0.96	1.00
	AM-3	0.08	0.09	0.22	0.38	0.53	0.70	0.83	0.88	0.96	1.00
	AM-4	0.08	0.09	0.15	0.27	0.44	0.39	0.57	0.68	0.80	1.00
	FT	0.02	0.43	0.84	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.05	0.40	0.75	0.93	0.99	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.03	0.21	0.39	0.71	0.82	0.85	0.95	0.98	0.98	1.00
	NN-DS	0.02	0.11	0.12	0.16	0.28	0.23	0.35	0.36	0.33	0.55
	NN-IS	0.07	0.02	0.10	0.11	0.10	0.15	0.13	0.11	0.10	0.01
$r^* = 4$	NN-OS	0.02	0.06	0.03	0.03	0.05	0.06	0.06	0.04	0.05	0.04
	AM-1	0.07	0.12	0.28	0.49	0.63	0.64	0.78	0.88	0.90	0.96
	AM-2	0.10	0.15	0.22	0.48	0.56	0.71	0.78	0.88	0.90	0.99
	AM-3	0.06	0.13	0.21	0.41	0.55	0.64	0.74	0.81	0.82	0.99
	AM-4	0.05	0.08	0.20	0.30	0.33	0.52	0.53	0.61	0.73	0.97

Table 5.4: Simulations for Inference for Matrix Completion

	SNR	0.000	0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000	2.000
	FT	0.08	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.06	0.80	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.03	0.20	0.53	0.74	0.81	0.97	0.96	0.97	1.00	1.00
	NN-DS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NN-IS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r^* = 1$	NN-OS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AM-1	0.07	0.17	0.51	0.75	0.91	0.99	0.99	0.99	1.00	1.00
	AM-2	0.09	0.12	0.39	0.62	0.77	0.90	0.93	0.97	1.00	1.00
	AM-3	0.07	0.13	0.21	0.38	0.60	0.70	0.79	0.87	0.90	1.00
	AM-4	0.06	0.05	0.15	0.21	0.30	0.41	0.55	0.69	0.66	0.95
	FT	0.08	0.65	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.08	0.61	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.05	0.16	0.36	0.54	0.68	0.81	0.84	0.87	0.93	1.00
	NN-DS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NN-IS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r^* = 2$	NN-OS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AM-1	0.09	0.16	0.29	0.56	0.73	0.79	0.84	0.92	0.93	0.99
	AM-2	0.05	0.14	0.19	0.36	0.64	0.77	0.86	0.91	0.92	1.00
	AM-3	0.08	0.05	0.18	0.31	0.46	0.62	0.73	0.77	0.88	0.98
	AM-4	0.08	0.09	0.15	0.11	0.20	0.25	0.43	0.48	0.52	0.94
	FT	0.08	0.52	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.06	0.44	0.87	0.97	1.00	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.07	0.11	0.23	0.37	0.51	0.64	0.75	0.85	0.90	0.98
	NN-DS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NN-IS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r^* = 3$	NN-OS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AM-1	0.06	0.14	0.22	0.29	0.52	0.64	0.70	0.73	0.80	0.95
	AM-2	0.04	0.08	0.18	0.29	0.47	0.55	0.73	0.74	0.81	1.00
	AM-3	0.08	0.08	0.09	0.22	0.28	0.40	0.49	0.65	0.72	0.98
	AM-4	0.09	0.06	0.09	0.12	0.14	0.24	0.25	0.36	0.42	0.86
	FT	0.08	0.36	0.79	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	LS	0.06	0.31	0.67	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	NN-OR	0.09	0.17	0.23	0.36	0.42	0.48	0.61	0.63	0.73	0.96
	NN-DS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NN-IS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r^* = 4$	NN-OS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AM-1	0.05	0.10	0.18	0.25	0.35	0.48	0.54	0.60	0.65	0.82
	AM-2	0.08	0.09	0.18	0.24	0.31	0.44	0.48	0.58	0.69	0.92
	AM-3	0.07	0.07	0.09	0.19	0.22	0.29	0.39	0.48	0.60	0.86
	AM-4	0.05	0.04	0.09	0.07	0.13	0.15	0.30	0.31	0.40	0.75

APPENDIX 1

Appendix of Chapter 2

This is the Appendix to Chapter 2.

A.2.1 Additional Simulation Results

In this section, we include additional results for the simulations of Section 2.5.

Table A.2.1: Simulations for β with Gaussian design and errors when $q = 3$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.924	0.886	0.922	0.936	0.910	0.872	0.942	0.946
	SILM	0.936	0.894	0.960	0.964	0.940	0.788	0.896	0.874
	EW_I	0.944	0.886	0.956	0.950	0.952	0.800	0.976	0.978
	EW_{II}	0.978	0.942	0.978	0.976	0.976	0.908	0.992	0.994
	EW_{III}	0.992	0.964	0.988	0.990	0.990	0.952	0.998	0.994

Table A.2.2: Simulations for β with Gaussian design and errors when $q = 1$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.946	0.878	0.938	0.946	0.934	0.900	0.948	0.944
	DLA	0.928	0.880	0.934	0.946	0.904	0.856	0.352	0.238
	SILM	0.932	0.872	0.936	0.956	0.918	0.858	0.130	0.034
	DML	1.000	0.996	0.998	0.994	0.990	0.982	1.000	1.000
	EW_I	0.830	0.768	0.922	0.932	0.932	0.862	0.976	0.984
	EW_{II}	0.866	0.810	0.952	0.962	0.956	0.900	0.984	0.992
	EW_{III}	0.904	0.852	0.974	0.976	0.964	0.922	0.990	0.998
AvgLen	LS	0.428	0.463	0.591	0.688	0.429	0.466	0.932	1.450
	DLA	0.502	0.539	0.693	0.699	0.539	0.555	0.548	0.506
	SILM	0.549	0.579	0.683	0.709	0.647	0.636	0.673	0.640
	DML	1.190	1.180	1.180	1.160	2.800	1.670	17.300	17.200
	EW_I	0.640	0.655	0.711	0.717	1.080	0.805	1.910	1.870
	EW_{II}	0.696	0.717	0.778	0.801	1.180	0.883	2.130	2.120
	EW_{III}	0.746	0.773	0.839	0.877	1.270	0.953	2.320	2.350

Table A.2.3: Simulations for β with Gaussian design and errors when $q = 3$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.914	0.882	0.928	0.938	0.950	0.848	0.924	0.948
	SILM	0.862	0.684	0.886	0.916	0.892	0.644	0.020	0.004
	EW_I	0.522	0.636	0.766	0.838	0.832	0.736	0.858	0.928
	EW_{II}	0.568	0.696	0.828	0.886	0.876	0.790	0.908	0.968
	EW_{III}	0.630	0.746	0.858	0.938	0.900	0.828	0.952	0.978

Table A.2.4: Simulations for β with double exponential design and errors when $q = 1$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.942	0.900	0.966	0.954	0.950	0.892	0.940	0.946
	DLA	0.960	0.876	0.974	0.964	0.950	0.872	0.154	0.102
	SILM	0.954	0.876	0.968	0.960	0.954	0.838	0.892	0.868
	DML	0.960	0.878	0.930	0.914	0.980	0.848	1.000	1.000
	EW$_I$	0.950	0.858	0.962	0.962	0.956	0.866	0.960	0.954
	EW$_{II}$	0.972	0.918	0.978	0.976	0.976	0.914	0.970	0.976
	EW$_{III}$	0.980	0.938	0.990	0.984	0.990	0.950	0.980	0.990
AvgLen	LS	0.456	0.492	0.683	0.829	0.432	0.466	0.910	1.450
	DLA	0.534	0.565	0.813	0.821	0.530	0.545	0.520	0.490
	SILM	0.574	0.596	0.785	0.825	0.619	0.603	0.611	0.592
	DML	0.756	0.702	0.874	0.875	1.480	0.892	12.700	13.300
	EW$_I$	0.716	0.691	0.856	0.877	1.070	0.798	2.000	1.910
	EW$_{II}$	0.792	0.772	0.932	0.974	1.180	0.891	2.180	2.120
	EW$_{III}$	0.860	0.844	1.000	1.060	1.280	0.973	2.330	2.300

Table A.2.5: Simulations for β with double exponential design and errors when $q = 3$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.928	0.834	0.904	0.930	0.940	0.884	0.926	0.940
	SILM	0.954	0.856	0.956	0.966	0.950	0.782	0.874	0.858
	EW$_I$	0.958	0.850	0.954	0.958	0.946	0.794	0.966	0.976
	EW$_{II}$	0.984	0.936	0.990	0.978	0.972	0.886	0.990	0.992
	EW$_{III}$	0.992	0.966	0.992	0.992	0.984	0.932	0.990	0.996

Table A.2.6: Simulations for β with double exponential design and errors when $q = 1$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.934	0.902	0.946	0.930	0.940	0.914	0.934	0.954
	DLA	0.940	0.884	0.920	0.918	0.910	0.876	0.332	0.280
	SILM	0.946	0.892	0.910	0.904	0.922	0.842	0.082	0.014
	DML	0.998	1.000	0.994	0.994	0.986	0.956	1.000	1.000
	EW_I	0.894	0.818	0.910	0.942	0.934	0.860	0.958	0.974
	EW_{II}	0.920	0.856	0.938	0.968	0.958	0.880	0.972	0.986
	EW_{III}	0.942	0.886	0.956	0.974	0.974	0.902	0.988	0.990
AvgLen	LS	0.454	0.495	0.677	0.831	0.434	0.468	0.899	1.470
	DLA	0.541	0.573	0.803	0.829	0.545	0.550	0.519	0.491
	SILM	0.602	0.629	0.788	0.842	0.652	0.627	0.603	0.586
	DML	1.360	1.270	1.480	1.460	2.830	1.660	22.200	22.000
	EW_I	0.726	0.715	0.854	0.892	1.100	0.806	2.020	1.910
	EW_{II}	0.791	0.789	0.941	1.000	1.190	0.884	2.250	2.170
	EW_{III}	0.851	0.855	1.020	1.100	1.280	0.953	2.460	2.390

Table A.2.7: Simulations for β with double exponential design and errors when $q = 3$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.936	0.878	0.950	0.940	0.948	0.854	0.934	0.942
	SILM	0.864	0.682	0.876	0.906	0.878	0.624	0.010	0.002
	EW_I	0.582	0.692	0.818	0.844	0.828	0.714	0.868	0.950
	EW_{II}	0.658	0.760	0.870	0.904	0.880	0.774	0.920	0.976
	EW_{III}	0.714	0.822	0.914	0.936	0.912	0.820	0.960	0.986

Table A.2.8: Simulations for β with scaled t design and errors when $q = 1$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.958	0.910	0.942	0.938	0.956	0.904	0.950	0.960
	DLA	0.954	0.878	0.962	0.948	0.946	0.878	0.198	0.114
	SILM	0.968	0.882	0.968	0.960	0.946	0.838	0.866	0.834
	DML	0.980	0.868	0.920	0.880	0.976	0.822	0.998	0.998
	EW$_I$	0.950	0.846	0.956	0.956	0.966	0.816	0.968	0.974
	EW$_{II}$	0.972	0.902	0.984	0.974	0.976	0.880	0.986	0.982
	EW$_{III}$	0.988	0.940	0.984	0.980	0.982	0.904	0.990	0.994
AvgLen	LS	0.490	0.515	0.750	0.933	0.453	0.479	0.951	1.600
	DLA	0.559	0.591	0.886	0.892	0.525	0.547	0.569	0.544
	SILM	0.611	0.620	0.884	0.928	0.618	0.607	0.687	0.672
	DML	0.819	0.751	1.140	1.170	1.410	0.902	13.200	14.400
	EW$_I$	0.806	0.739	0.967	0.982	1.100	0.817	2.240	2.110
	EW$_{II}$	0.882	0.828	1.060	1.090	1.210	0.914	2.450	2.330
	EW$_{III}$	0.952	0.907	1.140	1.180	1.300	0.999	2.640	2.530

Table A.2.9: Simulations for β with scaled t design and errors when $q = 3$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.922	0.872	0.940	0.944	0.936	0.856	0.926	0.936
	SILM	0.954	0.866	0.964	0.972	0.954	0.822	0.846	0.796
	EW$_I$	0.958	0.832	0.958	0.968	0.950	0.798	0.962	0.972
	EW$_{II}$	0.980	0.924	0.978	0.990	0.986	0.886	0.986	0.994
	EW$_{III}$	0.990	0.958	0.988	0.994	0.990	0.916	0.996	0.996

Table A.2.10: Simulations for β with scaled t design and errors when $q = 1$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AvgCov	LS	0.938	0.894	0.926	0.948	0.946	0.894	0.952	0.952
	DLA	0.924	0.888	0.908	0.936	0.900	0.874	0.412	0.348
	SILM	0.886	0.934	0.900	0.914	0.908	0.842	0.112	0.046
	DML	0.998	1.000	0.992	0.998	0.984	0.978	1.000	1.000
	EW$_I$	0.882	0.790	0.940	0.954	0.926	0.816	0.976	0.978
	EW$_{II}$	0.920	0.836	0.964	0.978	0.956	0.868	0.982	0.990
	EW$_{III}$	0.942	0.868	0.972	0.990	0.968	0.890	0.988	0.994
AvgLen	LS	0.487	0.521	0.742	0.938	0.458	0.478	0.956	1.610
	DLA	0.565	0.608	0.892	0.910	0.549	0.558	0.573	0.548
	SILM	0.643	0.657	0.902	0.957	0.659	0.638	0.696	0.668
	DML	1.540	1.370	1.850	1.750	2.870	1.730	22.800	23.200
	EW$_I$	0.814	0.753	0.983	1.010	1.110	0.831	2.140	2.150
	EW$_{II}$	0.886	0.830	1.090	1.140	1.210	0.911	2.390	2.420
	EW$_{III}$	0.952	0.898	1.190	1.250	1.300	0.983	2.610	2.670

Table A.2.11: Simulations for β with scaled t design and errors when $q = 3$ and $\beta = 1$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
AveCov	LS	0.950	0.906	0.944	0.918	0.918	0.874	0.950	0.948
	SILM	0.894	0.704	0.890	0.874	0.874	0.746	0.052	0.008
	EW$_I$	0.638	0.644	0.820	0.832	0.840	0.694	0.926	0.940
	EW$_{II}$	0.682	0.716	0.904	0.906	0.890	0.746	0.962	0.984
	EW$_{III}$	0.724	0.766	0.944	0.952	0.908	0.792	0.974	0.994

Table A.2.12: Simulations for σ_μ^2 with $s_\gamma = 15$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.762	0.734	0.768	0.816	0.892	0.906
	CHIVE_0	0.134	0.492	0.152	0.464	0.228	0.460
	CHIVE_2	0.380	0.584	0.392	0.560	0.408	0.554
	CHIVE_4	0.514	0.676	0.554	0.638	0.540	0.674
	CHIVE_6	0.646	0.740	0.632	0.698	0.624	0.690
	EW_I	0.328	0.696	0.422	0.732	0.388	0.652
	EW_{II}	0.628	0.756	0.630	0.784	0.588	0.786
	EW_{III}	0.690	0.606	0.672	0.690	0.718	0.816
AvgLen	LS	1.450	1.440	1.500	1.850	1.990	3.080
	CHIVE_0	0.538	0.873	0.583	0.999	1.060	2.140
	CHIVE_2	1.410	1.550	1.420	1.670	1.810	2.700
	CHIVE_4	1.910	1.980	1.910	2.090	2.270	3.070
	CHIVE_6	2.310	2.320	2.300	2.440	2.630	3.380
	EW_I	1.200	1.370	1.260	1.640	1.720	2.830
	EW_{II}	1.280	1.340	1.320	1.630	1.790	2.820
	EW_{III}	1.230	1.250	1.270	1.560	1.770	2.780

Table A.2.13: Simulations for σ_ε^2 with $s_\gamma = 15$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.874	0.864	0.870	0.848	0.876	0.864
	SL	0.308	0.646	0.386	0.620	0.466	0.606
	RCV-SIS	0.004	0.238	0.006	0.256	0.012	0.254
	EW_I	0.514	0.630	0.554	0.650	0.532	0.648
	EW_{II}	0.026	0.362	0.042	0.358	0.058	0.376
	EW_{III}	0.000	0.110	0.006	0.092	0.002	0.126
AvgLen	LS	0.481	0.462	0.467	0.478	0.479	0.483
	SL	0.781	0.702	0.766	0.722	0.753	0.717
	RCV-SIS	1.030	0.711	1.030	0.746	1.210	0.724
	EW_I	0.613	0.498	0.600	0.515	0.589	0.507
	EW_{II}	0.676	0.536	0.660	0.553	0.644	0.541
	EW_{III}	0.793	0.605	0.770	0.623	0.746	0.604

A.2.2 Proofs

A.2.2.1 Proofs for Section 2.2.2

Lemma A2.1. Consider the models given in equations (2.2.1) and (2.2.2). Under assumptions (2.1), (2.2 *), and (2.3),

$$\frac{\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\varepsilon}}{\sigma_\eta \operatorname{tr}(\boldsymbol{\Sigma}_\varepsilon)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Proof. By the Spectral Theorem, there exists a unitary matrix $\boldsymbol{\Gamma}$ and a diagonal matrix D such that $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{\Gamma}\mathbf{D}\boldsymbol{\Gamma}^\mathsf{T}$. Since $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ are both Gaussian and independent, there exists Gaussian vectors $\boldsymbol{\zeta} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{I}_n)$ and $\boldsymbol{\xi} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{I}_n)$ such that

$$\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\varepsilon} \overset{\mathcal{L}}{=} \sigma_\eta \boldsymbol{\zeta}^\mathsf{T}\mathbf{D}^{1/2}\boldsymbol{\xi}.$$

Then, by the Lindeberg Central Limit Theorem, it follows that

$$\frac{\boldsymbol{\zeta}^\mathsf{T}\mathbf{D}^{1/2}\boldsymbol{\xi}}{\sqrt{\operatorname{tr}(\mathbf{D})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Noting that $\operatorname{tr}(\mathbf{D}) = \operatorname{tr}(\mathbf{\Sigma})$ finishes the proof. $\square$

Proof of Theorem 2.3. The proof follows by combining Lemmata 2.15, 2.16, and A2.1. $\square$

A.2.2.2 Proofs for Section 2.2.3

Similar to the setting where $q = 1$, we will proceed in a few stages.

Lemma A2.2. Consider the models given in equations (2.2.1) and (2.2.2). Under assumptions (2.4) – (2.6),

1.
$$\left\| \left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)^{\mathsf{T}} \left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right) \right\| = o_{\mathbb{P}}(\sqrt{n}).$$

2.
$$\left\| \left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)^{\mathsf{T}} \mathbf{H} \right\| = o_{\mathbb{P}}(\sqrt{n}).$$

3.
$$n \left(\left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)^{\mathsf{T}} \left(\mathbf{X} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right) \right)^{-1} \xrightarrow{\mathbb{P}} \mathbf{\Sigma}_H^{-1}.$$

Proof. Indeed, note that $\left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)^{\mathsf{T}} \left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)$ is a positive definite matrix. By q applications of Lemma 2.15, each diagonal element is $o_{\mathbb{P}}(\sqrt{n})$, which proves the first claim. For the second part, Lemma 2.15 again shows that each diagonal element is $o_{\mathbb{P}}(\sqrt{n})$. It is left to show that each off diagonal element is also $o_{\mathbb{P}}(\sqrt{n})$. By symmetry, it suffices to consider the $(1, 2)$ element of $\left(\mathbf{N} - \mathbf{Z}\hat{\mathbf{\Delta}}_{\text{EW}}\right)^{\mathsf{T}} \mathbf{H}$. For simplicity, we write $\boldsymbol{\nu}$ to denote the first column of $\mathbf{N}$, $\hat{\boldsymbol{\delta}}_{\text{EW}}$ to denote the first column of $\hat{\mathbf{\Delta}}_{\text{EW}}$, $w_{\mathbf{m}}$ to denote the exponential weights of $\hat{\boldsymbol{\delta}}_{\text{EW}}$, $\boldsymbol{\eta}$ to denote the first column of $\mathbf{H}$, and $\boldsymbol{\xi}$ to denote the second column of $\mathbf{H}$. Then, the $(1, 2)$ element can be expressed as

$$\begin{aligned}
\left(\boldsymbol{\nu} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{\text{EW}}\right)^{\mathsf{T}} \boldsymbol{\xi} &= \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}} \boldsymbol{\nu}^{\mathsf{T}} \mathbf{P}_{\mathbf{m}}^{\perp} \boldsymbol{\xi} - \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}} \boldsymbol{\eta}^{\mathsf{T}} \mathbf{P}_{\mathbf{m}} \boldsymbol{\xi} \\
&= \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}}^{\mathsf{T}} \mathbf{P}_{\mathbf{m}}^{\perp} \boldsymbol{\xi} - \frac{1}{2} \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}} \left\| \mathbf{P}_{\mathbf{m}} \left(\boldsymbol{\xi} + \boldsymbol{\eta}\right) \right\|_2^2 \\
&\quad + \frac{1}{2} \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}} \left\| \mathbf{P}_{\mathbf{m}} \boldsymbol{\xi} \right\|_2^2 + \frac{1}{2} \sum_{\mathbf{m} \in \mathcal{M}_u} w_{\mathbf{m}} \left\| \mathbf{P}_{\mathbf{m}} \boldsymbol{\eta} \right\|_2^2 .
\end{aligned}$$

Applying Lemma 2.13 and Corollary 2.6.1 proves the second claim. Finally, note that

$$\begin{aligned}
&\left\|\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)-n\boldsymbol{\Sigma}_H\right\|_2 \\
&\quad\leq \left\|\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)\right\|_2+2\left\|\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\mathbf{H}\right\|_2 \\
&\qquad+\left\|\mathbf{H}^{\mathsf{T}}\mathbf{H}-n\boldsymbol{\Sigma}_H\right\|_2.
\end{aligned}$$

We have already shown that the first two terms are $o_{\mathbb{P}}(\sqrt{n})$. For the last term, by the Law of Large Numbers, it follows that

$$\left\|\mathbf{H}^{\mathsf{T}}\mathbf{H}-n\boldsymbol{\Sigma}_H\right\|_2=o_{\mathbb{P}}(n).$$

Therefore,

$$\frac{1}{n}\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)\xrightarrow{\mathbb{P}}\boldsymbol{\Sigma}_H.$$

Since $\boldsymbol{\Sigma}_H$ is assumed to be invertible, applying the Continuous Mapping Theorem finishes the proof. □

Proof of Theorem 2.4. For convenience, define the following matrices

$$\begin{aligned}
\mathbf{A}&\triangleq\left(\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)\right),\\
\mathbf{B}&\triangleq\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\mathbf{H},\\
\mathbf{C}&\triangleq\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{N}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right).
\end{aligned}$$

Applying Lemma 2.15 to each row separately, we see that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{EW}}&=\sqrt{n}\mathbf{A}^{-1}\left(\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\Delta}}_{\text{EW}}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{EW}}\right)\\
&=\sqrt{n}\mathbf{A}^{-1}\left(\mathbf{H}^{\mathsf{T}}\mathbf{H}\boldsymbol{\beta}+\mathbf{H}^{\mathsf{T}}\boldsymbol{\varepsilon}+\mathbf{R}\right),
\end{aligned}$$

where $\|\mathbf{R}\|_1=o_{\mathbb{P}}(\sqrt{n})$. But, from Lemma A2.2, we have that

$$n\left\|\mathbf{A}^{-1}\right\|_2\xrightarrow{\mathbb{P}}\left\|\boldsymbol{\Sigma}_H^{-1}\right\|_2,$$

which is finite since $\mathbf{\Sigma}_H$ is invertible by assumption. Therefore,

$$\left\|\sqrt{n}\mathbf{A}^{-1}\mathbf{R}\right\| \leq \left(n\left\|\mathbf{A}^{-1}\right\|_2\right)\left(n^{-1/2}\left\|\mathbf{R}\right\|_1\right) \xrightarrow{\mathbb{P}} 0.$$

Now, note that

$$\mathbf{H}^\mathsf{T}\mathbf{H} = \mathbf{A} - \mathbf{B} - \mathbf{B}^\mathsf{T} - \mathbf{C}.$$

Hence,

$$\sqrt{n}\mathbf{A}^{-1}\mathbf{H}^\mathsf{T}\mathbf{H}\boldsymbol{\beta} = \sqrt{n}\boldsymbol{\beta} - \sqrt{n}\mathbf{A}^{-1}\left(\mathbf{B} + \mathbf{B}^\mathsf{T} + \mathbf{C}\right)\boldsymbol{\beta}.$$

Again, by Lemma A2.2,

$$\left\|\sqrt{n}\mathbf{A}^{-1}\left(\mathbf{B} + \mathbf{B}^\mathsf{T} + \mathbf{C}\right)\boldsymbol{\beta}\right\|_2 \leq \left(n\left\|\mathbf{A}^{-1}\right\|_2\right)\left(n^{-1/2}\left\|\mathbf{B} + \mathbf{B}^\mathsf{T} + \mathbf{C}\right\|_2\right)\left\|\boldsymbol{\beta}\right\|_2 \xrightarrow{\mathbb{P}} 0.$$

Finally, by the Multivariate Central Limit Theorem,

$$n^{-1/2}\mathbf{H}^\mathsf{T}\boldsymbol{\varepsilon} \xrightarrow{\mathcal{L}} \mathcal{N}_q\left(\mathbf{0}_q, \sigma_\varepsilon^2\mathbf{\Sigma}_H\right).$$

Since $n\mathbf{A}^{-1} \xrightarrow{\mathbb{P}} \mathbf{\Sigma}_H^{-1}$, it follows by Slutsky's Theorem that

$$\sqrt{n}\mathbf{A}^{-1}\mathbf{H}^\mathsf{T}\boldsymbol{\varepsilon} \xrightarrow{\mathcal{L}} \mathcal{N}_q\left(\mathbf{0}_q, \sigma_\varepsilon^2\mathbf{\Sigma}_H^{-1}\right),$$

which finishes the proof. □

Proof of Proposition 2.5. This follows from Lemma A2.2. □

A.2.2.3 Proofs for Section 2.3.1

Proof of Proposition 2.7. Letting $\hat{\boldsymbol{\gamma}}$ denote the least-squares estimator for $\boldsymbol{\gamma}$, it is known that $\hat{\boldsymbol{\gamma}}$ is efficient for estimating $\boldsymbol{\gamma}$ in the low-dimensional linear model. Since $\mathbf{Z}$ is assumed to be of full rank, there exists a smooth re-parameterization of the problem given by $(\boldsymbol{\gamma}, \sigma_\varepsilon^2) \mapsto \left(\sigma_\mu^2, \boldsymbol{\vartheta}, \sigma_\varepsilon^2\right)$, where $\left(\sigma_\mu^2, \boldsymbol{\vartheta}\right)$ is the polar representation of $\left\|\mathbf{Z}_{\boldsymbol{S}_\gamma}\boldsymbol{\gamma}\right\|_2^2$. Taking the bowl-shaped loss to be quadratic in the first component, the result follows from the arguments of Section 2.3 of Bickel et al. (1993) since $\left\|\mathbf{P}_{\boldsymbol{S}_\gamma}\mathbf{y}\right\|_2^2 = \left\|\mathbf{Z}\hat{\boldsymbol{\gamma}}\right\|_2^2$. □

The proof for Theorem 2.8 will rely on the proof of Theorem 2.10 from Section A.2.2.4.

Proof of Theorem 2.8. Indeed, we may write

$$\frac{1}{n}\|\mathbf{y}\|_2^2 = \frac{1}{n}\|\boldsymbol{\mu}\|_2^2 + \frac{2}{n}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\varepsilon} + \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2$$

Note that, from equations (A.2.2.3), (A.2.2.4) and (A.2.2.5), it follows that

$$\begin{aligned}
\hat{\sigma}^2_{\mu,I} &= \frac{1}{n}\|\boldsymbol{\mu}\|_2^2 + \frac{2}{n}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\varepsilon} + o_{\mathbb{P}}\left(n^{-1/2}\right), \\
\hat{\sigma}^2_{\mu,II} &= \frac{1}{n}\|\boldsymbol{\mu}\|_2^2 + \frac{2}{n}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\varepsilon} + o_{\mathbb{P}}\left(n^{-1/2}\right), \\
\hat{\sigma}^2_{\mu,III} &= \frac{1}{n}\|\boldsymbol{\mu}\|_2^2 + \frac{2}{n}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\varepsilon} + o_{\mathbb{P}}\left(n^{-1/2}\right).
\end{aligned}$$

By the Multivariate Central Limit Theorem, it follows that

$$\sqrt{n}\begin{pmatrix} n^{-1}\|\boldsymbol{\mu}\|^2 - \sigma^2_\mu \\ 2n^{-1}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\varepsilon} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \kappa & 0 \\ 0 & 4\sigma^2_\varepsilon\sigma^2_\mu \end{pmatrix}\right).$$

Applying the Cramér-Wold device finishes the proof. □

Proof of Proposition 2.9. Indeed,

$$\begin{aligned}
\hat{\kappa}_\mu =& \frac{1}{n}\sum_{j=1}^n \left(\left(\boldsymbol{\mu}_j^2 - \sigma^2_\mu\right) + \left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)^2 + 2\left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j - \left(\hat{\sigma}^2_\mu - \sigma^2_\mu\right)\right)^2 \\
=& \frac{1}{n}\sum_{j=1}^n \left(\boldsymbol{\mu}_j^2 - \sigma^2_\mu\right)^2 + \frac{1}{n}\sum_{j=1}^n \left(\left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)^2 + 2\left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j - \left(\hat{\sigma}^2_\mu - \sigma^2_\mu\right)\right)^2 \\
&+ \frac{2}{n}\sum_{j=1}^n \left(\boldsymbol{\mu}_j^2 - \sigma^2_\mu\right)\left(\left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)^2 + 2\left(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j - \left(\hat{\sigma}^2_\mu - \sigma^2_\mu\right)\right).
\end{aligned}$$

Applying the Law of Large Numbers yields

$$\frac{1}{n}\sum_{j=1}^n \left(\boldsymbol{\mu}_j^2 - \sigma^2_\mu\right)^2 \xrightarrow{\mathbb{P}} \kappa_\mu. \tag{A.2.2.1}$$

By the triangle inequality and Cauchy-Schwarz, it follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\left(\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2+2\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j-\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)\right)^2\\
&\qquad\le\frac{4}{n}\left\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right\|_4^4+\frac{8}{n}\sum_{j=1}^{n}\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2\boldsymbol{\mu}_j^2+4\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)^2\\
&\qquad\le\frac{4}{n}\left\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right\|_2^4+\frac{8}{n}\left\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right\|_2^2\left\|\boldsymbol{\mu}\right\|_2^2{}_4+4\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)^2.
\end{aligned}$$

From Theorem 2.8, we see that $\hat{\sigma}_\mu^2\xrightarrow{\mathbb{P}}\sigma_\mu^2$. Therefore, combining this with Proposition 2.1 shows that

$$\frac{1}{n}\sum_{j=1}^{n}\left(\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2+2\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j-\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)\right)^2\xrightarrow{\mathbb{P}}0. \tag{A.2.2.2}$$

Now, by another application of Cauchy-Schwarz,

$$\begin{aligned}
&\frac{2}{n}\sum_{j=1}^{n}\left|\left(\boldsymbol{\mu}_j^2-\sigma_\mu^2\right)\left(\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2+2\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j-\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)\right)\right|\\
&\quad\le\frac{2}{n}\left(\sum_{j=1}^{n}\left(\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2+2\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j-\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)\right)^2\right)^{1/2}\\
&\qquad\times\left(\sum_{j=1}^{n}\left(\boldsymbol{\mu}_j^2-\sigma_\mu^2\right)^2\right)^{1/2}.
\end{aligned}$$

From equations (A.2.2.1) and (A.2.2.2), it will follow that

$$\frac{2}{n}\sum_{j=1}^{n}\left|\left(\boldsymbol{\mu}_j^2-\sigma_\mu^2\right)\left(\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)^2+2\left(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}_j\right)\boldsymbol{\mu}_j-\left(\hat{\sigma}_\mu^2-\sigma_\mu^2\right)\right)\right|\xrightarrow{\mathbb{P}}0.$$

Combining the results finishes the proof. □

A.2.2.4 Proofs for Section 2.3.2

Proof of Theorem 2.10. Indeed, note that

$$\hat{\sigma}_{\varepsilon,I}^2=\frac{1}{n}\left(\left\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right\|_2^2+\varepsilon^{\mathsf{T}}\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)+\left\|\boldsymbol{\varepsilon}\right\|_2^2\right)=\frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|_2^2+o_{\mathbb{P}}\left(n^{-1/2}\right), \tag{A.2.2.3}$$

where the last equality follows from Proposition 2.1 and Lemma 2.15. Next, some algebra shows that $\hat{\sigma}^2_{\varepsilon,III}$ may be decomposed as

$$\begin{aligned}\hat{\sigma}^2_{\varepsilon,III} =& \frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|_2^2 + \frac{1}{n}\sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{m},y}\left(\left\|\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\mu}\right\|^2 + 2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\varepsilon}\right)\\ &- \frac{1}{n}\sum_{\mathbf{k}\in\mathcal{M}_{u_\gamma}}\sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{k},y}w_{\mathbf{m},y}\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}^{\perp}_{\mathbf{k}}\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\mu} + \boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{P}_{\mathbf{k}}\mathbf{P}_{\mathbf{m}}\boldsymbol{\varepsilon}\right).\end{aligned}$$

Applying Corollary 2.6.1 yields

$$\frac{1}{n}\sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{m},y}\left(\left\|\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\mu}\right\|_2^2 + 2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\varepsilon}\right) = o_{\mathbb{P}}\left(n^{-1/2}\right).$$

For the other term, it follows from Cauchy-Schwarz, Lemma 2.13, and Corollary 2.6.1 that

$$\frac{1}{n}\sum_{\mathbf{k}\in\mathcal{M}_{u_\gamma}}\sum_{\mathbf{m}\in\mathcal{M}_{u_\gamma}} w_{\mathbf{k},y}w_{\mathbf{m},y}\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{P}^{\perp}_{\mathbf{k}}\mathbf{P}^{\perp}_{\mathbf{m}}\boldsymbol{\mu} + \boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{P}_{\mathbf{k}}\mathbf{P}_{\mathbf{m}}\boldsymbol{\varepsilon}\right) = o_{\mathbb{P}}\left(n^{-1/2}\right).$$

Thus, this implies that

$$\hat{\sigma}^2_{\varepsilon,III} = \frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|_2^2 + o_{\mathbb{P}}\left(n^{-1/2}\right). \tag{A.2.2.4}$$

Now, by Jensen's inequality,

$$\hat{\sigma}^2_{\varepsilon,I} \leq \hat{\sigma}^2_{\varepsilon,II} \leq \hat{\sigma}^2_{\varepsilon,III}.$$

Therefore, we may conclude that

$$\hat{\sigma}^2_{\varepsilon,II} = \frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|_2^2 + o_{\mathbb{P}}\left(n^{-1/2}\right). \tag{A.2.2.5}$$

The asymptotic distribution for all three estimators follows by applying the Central Limit Theorem, which finishes the proof. □

Proof of Corollary 2.3.2. Indeed,

$$\hat{\sigma}^2_{\varepsilon,I} = \frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|_2^2 + o_{\mathbb{P}}\left(n^{-1/2}\right).$$

From the proof of Lemma A2.1, we may apply the Spectral Theorem to obtain the following decomposition: $\boldsymbol{\Sigma}_{\varepsilon} = \boldsymbol{\Gamma}\mathbf{D}\boldsymbol{\Gamma}^{\mathsf{T}}$. Then, for $\boldsymbol{\xi} \sim \mathcal{N}_n\left(\mathbf{0}_n, \mathbf{I}_n\right)$, it follows that $\mathbf{D}^{1/2}\boldsymbol{\xi} \overset{\mathcal{L}}{=} \boldsymbol{\varepsilon}$. A direct variance calculation for $n^{-1}\left\|\mathbf{D}^{1/2}\boldsymbol{\xi}\right\|_2^2$ finishes the proof. □

Proof of Proposition 2.11. The proof is similar to the proof of Proposition 2.9. □

APPENDIX 2

Appendix of Chapter 3

A.3.1 Proofs

A.3.1.1 Proofs for Section 5.3

We start with a simple lemma, which is a consequence of Lemma 6.3 of Law and Ritov (2021b).

Lemma A3.1. Consider the model given in equation (3.1.1). Assume (3.1), (3.2), (3.3), (3.4), and (3.5). Then,

$$\begin{aligned}
&\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 = \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 + o_{\mathbb{P}}(n^{\tau}), \\
&\|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 = \|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(n^{\tau}).
\end{aligned}$$

Proof. Indeed, we may expand the left hand side to obtain

$$\begin{aligned}
\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}}) = {} & \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 + 2(\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}) \\
& + \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2.
\end{aligned}$$

By assumptions (3.2) and (3.3), it follows that $\mathbf{P}^{\perp}_{\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\varepsilon) \sim \mathcal{SG}(K_{\nu}\,\lambda_{\text{max}}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}) + K_{\varepsilon})$. Thus, applying both parts of Lemma 6.3 of Law and Ritov (2021b) yields

$$\begin{aligned}
&\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 = o_{\mathbb{P}}(n^{\tau}), \\
&2(\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon}) = o_{\mathbb{P}}(n^{\tau}).
\end{aligned}$$

This proves the first claim. The proof for the other claim is identical and is omitted. $\square$

Proof of Theorem 3.1. Indeed, expanding the numerator of F_{ld}, we have that

$$r^{-1}_{\mathbf{Z}\ominus(\mathbf{X}_{\mathcal{S}},\mathbf{W})}\|\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\mathcal{S}},\mathbf{W})}\mathbf{y}\|_2^2 = r^{-1}_{\mathbf{Z}\ominus(\mathbf{X}_{\mathcal{S}},\mathbf{W})}\|\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\mathcal{S}},\mathbf{W})}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2.$$

For F_{EW}, using Lemma A3.1, we may similarly expand the numerator to obtain

$$r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 = r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 + o_{\mathbb{P}}(n^{\tau-1}).$$

Since $\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})\subseteq\mathbf{Z}\ominus\mathbf{W}$, by Pythagoras, we have that

$$r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 = r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 + r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}-\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})})(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2.$$

Note that $\operatorname{rank}(\mathbf{X}_{\boldsymbol{S}})=s$, so $\operatorname{rank}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}-\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}) = r_{\mathbf{Z}\ominus\mathbf{W}} - r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}\leq s$, implying that $r_{\mathbf{Z}\ominus\mathbf{W}}\asymp r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}$ by assumption (3.5). Therefore,

$$r_{\mathbf{Z}\ominus\mathbf{W}}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 = r_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}^{-1}\|\mathbf{P}_{\mathbf{Z}\ominus(\mathbf{X}_{\boldsymbol{S}},\mathbf{W})}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 + o_{\mathbb{P}}(n^{\tau-1}).$$

By a similar argument, it can be shown for the denominators that

$$r_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{X}_{\boldsymbol{S}},\mathbf{Z},\mathbf{W})}^{\perp}\mathbf{y}\|_2^2 = r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 + o_{\mathbb{P}}(n^{\tau-1}).$$

Dividing the last two displays finishes the proof. □

Proof of Theorem 3.2. We start by recalling the definition of $\varphi_{F,\delta}$,

$$\varphi_{F,\delta} = \mathbb{1}\left(F_{\text{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp},\delta}\right).$$

Applying Lemma A3.1, we may lower bound F_{EW} by

$$\begin{aligned}
F_{\text{EW}} &= \frac{\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2/r_{\mathbf{Z}\ominus\mathbf{W}}}{\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2^2/r_{(\mathbf{Z},\mathbf{W})^\perp}} + o_{\mathbb{P}}(n^{\tau-1}) \\
&\geq \frac{\lambda_{\max,r_{\mathbf{Z}\ominus\mathbf{W}}}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}) + \sigma_\varepsilon^2}{\sigma_\varepsilon^2} F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp}} + o_{\mathbb{P}}(n^{\tau-1}) \\
&\geq \frac{hn^{\tau-1}+\sigma_\varepsilon^2}{\sigma_\varepsilon^2} F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp}} + o_{\mathbb{P}}(n^{\tau-1}),
\end{aligned}$$

We would like to show that, for n sufficiently large,

$$\mathbb{P}_{H_0}\left(F_{\text{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp},\delta}\right) + \mathbb{P}_{H_1}\left(F_{\text{EW}} \leq F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp},\delta}\right) < 1.$$

It suffices to show that

$$\mathbb{P}_{H_1}\left(F_{\text{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}},r_{(\mathbf{Z},\mathbf{W})^\perp},\delta}\right) > \delta.$$

We start by providing an upper bound on $F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}, \delta}$. From Lemma 1 of Laurent and Massart (2000), it follows that, for a χ^2_d random variable,

$$\mathbb{P}\left(\chi^2_d > d + 2\sqrt{dx} + 2x\right) \leq \exp(-x),$$
$$\mathbb{P}\left(\chi^2_d \leq d - 2\sqrt{dy}\right) \leq \exp(-y).$$

Therefore, it follows that, for any $x, y > 0$,

$$\mathbb{P}\left(F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}} > \frac{1 + 2\sqrt{x/r_{\mathbf{Z}\ominus\mathbf{W}}} + 2x/r_{\mathbf{Z}\ominus\mathbf{W}}}{1 - 2\sqrt{y/r_{(\mathbf{Z},\mathbf{W})^{\perp}}}}\right) \leq \exp(-x) + \exp(-y).$$

By choosing $x, y > 0$ such that

$$\exp(-x) + \exp(-y) \leq \delta,$$

then

$$F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}, \delta} \leq \frac{1 + 2\sqrt{x/r_{\mathbf{Z}\ominus\mathbf{W}}} + 2x/r_{\mathbf{Z}\ominus\mathbf{W}}}{1 - 2\sqrt{y/r_{(\mathbf{Z},\mathbf{W})^{\perp}}}} = 1 + \mathcal{O}(n^{-1/2}).$$

Let $a, b > 0$ be constants that will be chosen later. Define the event $\mathscr{T}$ as

$$\mathscr{T} \triangleq \left\{\chi^2_{r_{\mathbf{Z}\ominus\mathbf{W}}} > r_{\mathbf{Z}\ominus\mathbf{W}} - 2\sqrt{r_{\mathbf{Z}\ominus\mathbf{W}}a}, \chi^2_{r_{(\mathbf{Z},\mathbf{W})^{\perp}}} \leq r_{(\mathbf{Z},\mathbf{W})^{\perp}} + 2\sqrt{r_{(\mathbf{Z},\mathbf{W})^{\perp}}b} + 2b\right\}.$$

Again, by Lemma 1 of Laurent and Massart (2000),

$$\mathbb{P}\left(\mathscr{T}^{\mathsf{c}}\right) \leq \exp\left(-a\right) + \exp\left(-b\right).$$

On $\mathscr{T}$, it follows that

$$F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}} = 1 + \mathcal{O}(n^{-1/2}).$$

Now,

$$\begin{aligned}\mathbb{P}_{H_1}\left(F_{\mathrm{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}, \delta}\right) &\geq \mathbb{P}_{H_1}\left(F_{\mathrm{EW}} > F_{r_{\mathbf{Z}\ominus\mathbf{W}}, r_{(\mathbf{Z},\mathbf{W})^{\perp}}, \delta}, \mathscr{T}\right)\\ &\geq \mathbb{P}_{H_1}\left(\frac{hn^{\tau-1} + \sigma^2_{\varepsilon}}{\sigma^2_{\varepsilon}}\left(1 + \mathcal{O}(n^{-1/2})\right) + o_{\mathbb{P}}(n^{\tau-1}) > \left(1 + \mathcal{O}(n^{-1/2})\right), \mathscr{T}\right).\end{aligned}$$

Noting that $\tau - 1 \geq -1/2$ and letting h be sufficiently large, independent of n, finishes the proof. $\square$

Proof of Theorem 3.3. The proof is an immediate consequence of Theorem 3.5 by setting ν to follow a degenerate distribution at zero. $\square$

Before proving Theorem 3.4, we present an extension of Proposition 2.1 of Law and Ritov (2021b) to the setting of out of sample prediction for exponential weighting with Gaussian designs.

Consider a high-dimensional linear model given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.3.1.1}$$

with $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$, and $\|\boldsymbol{\beta}\|_0 = s_n = s$. Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \in \mathbb{R}^{n/2\times p}$ and $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\varepsilon}^{(1)}, \boldsymbol{\varepsilon}^{(2)} \in \mathbb{R}^{n/2}$ be the data obtained by data splitting into two equal halves. For $\mathbf{m} \in \mathcal{M}$, denote by $\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)} \triangleq (\mathbf{X}^{(1)\mathsf{T}}\mathbf{X}^{(1)})^{-}\mathbf{X}^{(1)\mathsf{T}}\mathbf{y}^{(1)}$. Then, the exponential weights are given by

$$\tilde{w}_{\mathbf{m}}^{(1)} \triangleq \frac{\exp\left(-\frac{1}{\alpha}\|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbf{m}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)}\|_2^2\right)}{\sum_{\mathbf{k}\in\mathcal{M}_u}\exp\left(-\frac{1}{\alpha}\|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbf{k}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{k}}^{(1)}\|_2^2\right)}.$$

Define $\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(2)}$ and $\tilde{w}_{\mathbf{m}}^{(2)}$ similarly. Then,

$$\tilde{\boldsymbol{\beta}}_{\text{EW}} \triangleq \sum_{\mathbf{m}\in\mathcal{M}}\left(\tilde{w}_{\mathbf{m}}^{(1)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(2)} + \tilde{w}_{\mathbf{m}}^{(2)}\tilde{\boldsymbol{\beta}}_{\mathbf{m}}^{(1)}\right).$$

Lemma A3.2. Consider the model given in equation (A.3.1.1). Assume that the rows of $\mathbf{X}$ are independent and identically distributed $\mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$ with $\max(\text{diag}(\Sigma_{\mathbf{X}})) = \mathcal{O}(1)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma_\varepsilon^2\mathbf{I}_n)$. Assume further that the chosen sequence of sparsities $u_n = u \geq s$ for n sufficiently large with $u = o(n^\tau/\log(p))$. If $\mathbf{x}_{\text{new}} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$ is independent of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ and $\alpha > 16\max(\text{diag}(\boldsymbol{\Sigma}_X), \sigma_\varepsilon^2)$, then

$$\mathbb{E}\left(\mathbf{x}_{\text{new}}^{\mathsf{T}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^2 = o(n^{\tau-1}).$$

where the expectation is over the joint distribution of $\mathbf{x}_{\text{new}}$, $\mathbf{X}$, and $\boldsymbol{\varepsilon}$.

Proof. By properties of the Gaussian distribution, for any $\mathbf{m} \in \mathcal{M}$, there exists vectors $\boldsymbol{\theta}_{\mathbf{m}} \in \mathbb{R}^u$ and $\boldsymbol{\xi}_{\mathbf{m}} \in \mathbb{R}^n$ such that

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\theta}_{\mathbf{m}} + \boldsymbol{\xi}_{\mathbf{m}}.$$

Here, the vector $\boldsymbol{\theta}_{\mathbf{m}}$ is a fixed vector of regression coefficients and $\boldsymbol{\xi}_{\mathbf{m}} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma_{\xi,\mathbf{m}}^2\mathbf{I}_n)$ is independent of $\mathbf{X}_{\mathbf{m}}$. Similarly, $\mathbf{x}_{\text{new}}^\mathsf{T}\boldsymbol{\beta} = \mathbf{x}_{\text{new},\mathbf{m}}^\mathsf{T}\boldsymbol{\theta}_{\mathbf{m}} + \boldsymbol{\xi}_{\text{new},\mathbf{m}}$. Then, by convexity, it follows that

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{x}_{\text{new}}^\mathsf{T}(\tilde{\boldsymbol{\beta}}_{\text{EW}} - \boldsymbol{\beta})\right)^2 \\
&\quad \le 2\mathbb{E}\left\{\sum_{\mathbf{m}\in\mathcal{M}} \tilde{w}_{\mathbf{m}}^{(1)}\left[\mathbf{x}_{\text{new}}^\mathsf{T}\left((\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\mathbf{X}_{\mathbf{m}}^{(2)}\boldsymbol{\theta}_{\mathbf{m}} + \boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)}) - \boldsymbol{\theta}_{\mathbf{m}}\right) - \boldsymbol{\xi}_{\text{new},\mathbf{m}}\right]\right\}^2 \\
&\qquad + 2\mathbb{E}\left\{\sum_{\mathbf{m}\in\mathcal{M}} \tilde{w}_{\mathbf{m}}^{(2)}\left[\mathbf{x}_{\text{new}}^\mathsf{T}\left((\mathbf{X}_{\mathbf{m}}^{(1)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(1)})^{-}\mathbf{X}_{\mathbf{m}}^{(1)\mathsf{T}}(\mathbf{X}_{\mathbf{m}}^{(1)}\boldsymbol{\theta}_{\mathbf{m}} + \boldsymbol{\xi}_{\mathbf{m}}^{(1)} + \boldsymbol{\varepsilon}^{(1)}) - \boldsymbol{\theta}_{\mathbf{m}}\right) - \boldsymbol{\xi}_{\text{new},\mathbf{m}}\right]\right\}^2.
\end{aligned}
$$

By symmetry, it suffices to only consider the first term. Note that

$$
\begin{aligned}
&\mathbb{E}\left\{\sum_{\mathbf{m}\in\mathcal{M}} \tilde{w}_{\mathbf{m}}^{(1)}\left[\mathbf{x}_{\text{new}}^\mathsf{T}\left((\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\mathbf{X}_{\mathbf{m}}^{(2)}\boldsymbol{\theta}_{\mathbf{m}} + \boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)}) - \boldsymbol{\theta}_{\mathbf{m}}\right) - \boldsymbol{\xi}_{\text{new},\mathbf{m}}\right]\right\}^2 \\
&\quad = \mathbb{E}\left\{\sum_{\mathbf{m}\in\mathcal{M}} \tilde{w}_{\mathbf{m}}^{(1)}\left[\mathbf{x}_{\text{new}}^\mathsf{T}(\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)}) - \boldsymbol{\xi}_{\text{new},\mathbf{m}}\right]\right\}^2 \\
&\quad \le 2\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left\{\tilde{w}_{\mathbf{m}}^{(1)}\left(\mathbf{x}_{\text{new}}^\mathsf{T}(\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)})\right)^2\right\} \\
&\qquad + 2\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right).
\end{aligned}
$$

Now, a direct calculation shows that

$$
\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left\{\tilde{w}_{\mathbf{m}}^{(1)}\left(\mathbf{x}_{\text{new}}^\mathsf{T}(\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)})\right)^2\right\} = \mathcal{O}\left(\frac{u}{n-u-1}\right).
$$

By assumption, since $u = o(n^\tau/\log(p))$, it follows that

$$
\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left\{\tilde{w}_{\mathbf{m}}^{(1)}\left(\mathbf{x}_{\text{new}}^\mathsf{T}(\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}\mathbf{X}_{\mathbf{m}}^{(2)})^{-}\mathbf{X}_{\mathbf{m}}^{(2)\mathsf{T}}(\boldsymbol{\xi}_{\mathbf{m}}^{(2)} + \boldsymbol{\varepsilon}^{(2)})\right)^2\right\} = o(n^{\tau-1}).
$$

Hence, it is left to show that

$$
\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right) = o(n^{\tau-1}).
$$

To this end, for an arbitrary fixed value of $t > 0$, define

$$
\mathcal{A}_t \triangleq \{\mathbf{m} \in \mathcal{M} : \sigma_{\xi,m}^2 \le tn^{\tau-1}\}.
$$

Since $\boldsymbol{\beta} = \boldsymbol{\beta}_{\boldsymbol{S}}$ with $\|\boldsymbol{\beta}\|_0 = s$ and $u \ge s$ for n sufficiently large, then $\boldsymbol{S} \in \mathcal{A}_t$ for n sufficiently large. Now,

$$\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right) = \sum_{\mathbf{m}\in\mathcal{A}_t} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right) + \sum_{\mathbf{m}\in\mathcal{A}_t} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right)$$
$$\leq tn^{\tau-1} + \sum_{\mathbf{m}\in\mathcal{A}_t} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)}\boldsymbol{\xi}_{\text{new},\mathbf{m}}^2\right).$$

For $\mathbf{m}\in\mathcal{M}$, write $\mathbf{P}_{\mathbf{m}}^{(1)}$ to denote the projection onto the column space of $\mathbf{X}_{\mathbf{m}}^{(1)}$. Then, for $\mathbf{m}\in\mathcal{A}_t^{\mathsf{c}}$,

$$\tilde{w}_{\mathbf{m}}^{(1)} \leq \exp\left(-\frac{1}{\alpha}\left(\|(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})(\boldsymbol{\xi}_{\mathbf{m}}^{(1)}+\boldsymbol{\varepsilon}^{(1)})\|_2^2 - \|(\mathbf{I}_{n/2}-\mathbf{P}_{\boldsymbol{S}}^{(1)})\boldsymbol{\varepsilon}^{(1)}\|_2^2\right)\right)$$
$$\leq \exp\left(-\frac{1}{\alpha}\left(\|(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\|_2^2 + 2{\boldsymbol{\varepsilon}^{(1)}}^{\mathsf{T}}(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)} + \|\mathbf{P}_{\boldsymbol{S}}^{(1)}\boldsymbol{\varepsilon}^{(1)}\|_2^2 - \|\mathbf{P}_{\mathbf{m}}^{(1)}\boldsymbol{\varepsilon}^{(1)}\|_2^2\right)\right)$$
$$\leq \exp\left(-\frac{1}{\alpha}\left(\|(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\|_2^2 + 2{\boldsymbol{\varepsilon}^{(1)}}^{\mathsf{T}}(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)} + \|\mathbf{P}_{\boldsymbol{S}}^{(1)}\boldsymbol{\varepsilon}^{(1)}\|_2^2\right)\right).$$

Now, by the Cauchy-Schwarz inequality, it follows that

$$\mathbb{E}\tilde{w}_{\mathbf{m}}^{(1)} \leq \left\{\mathbb{E}\exp\left[-\frac{2}{\alpha}\left(\|(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\|_2^2 + \|\mathbf{P}_{\boldsymbol{S}}^{(1)}\boldsymbol{\varepsilon}^{(1)}\|_2^2\right)\right]\right\}^{1/2}$$
$$\times\left\{\mathbb{E}\exp\left[-\frac{4}{\alpha}{\boldsymbol{\varepsilon}^{(1)}}^{\mathsf{T}}(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\right]\right\}^{1/2}.$$

A direct calculation yields,

$$\mathbb{E}\exp\left[-\frac{2}{\alpha}\left(\|(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\|_2^2 + \|\mathbf{P}_{\boldsymbol{S}}^{(1)}\boldsymbol{\varepsilon}^{(1)}\|_2^2\right)\right] \leq \left(1+\frac{4\sigma_{\xi,\mathbf{m}}^2}{\alpha}\right)^{-(n/2-u)/2}\left(1+\frac{4\sigma_{\varepsilon}^2}{\alpha}\right)^{-u/2},$$
$$\mathbb{E}\exp\left[-\frac{4}{\alpha}{\boldsymbol{\varepsilon}^{(1)}}^{\mathsf{T}}(\mathbf{I}_{n/2}-\mathbf{P}_{\mathbf{m}}^{(1)})\boldsymbol{\xi}_{\mathbf{m}}^{(1)}\right] \leq \left(1-\frac{16\sigma_{\xi,\mathbf{m}}^2\sigma_{\varepsilon}^2}{\alpha^2}\right)^{-(n/2-u)/2}.$$

Using the inequality

$$(1-2x)^{-1/2} \leq \exp(2x^2+x)$$

for $|x|<1/4$, it follows that

$$\mathbb{E}\tilde{w}_{\mathbf{m}}^{(1)} \leq \exp\left\{-\frac{\sigma_{\xi,\mathbf{m}}^2(n-2u)}{4\alpha}\left[1-\left(\frac{4(\sigma_{\xi,\mathbf{m}}^2+\sigma_{\varepsilon}^2)}{\alpha}+\frac{64\sigma_{\xi,\mathbf{m}}^2\sigma_{\varepsilon}^2}{\alpha^3}\right)\right] - \frac{\sigma_{\varepsilon}^2 u}{2\alpha}\left(1-\frac{8\sigma_{\varepsilon}^2}{\alpha}\right)\right\}.$$

By the choice of α,

$$1 - \left(\frac{4(\sigma_{\xi,\mathbf{m}}^2 + \sigma_\varepsilon^2)}{\alpha} + \frac{64\sigma_{\xi,\mathbf{m}}^2 \sigma_\varepsilon^2}{\alpha^3} \right) > 0,$$

then

$$\mathbb{E}\tilde{w}_{\mathbf{m}}^{(1)} \leq \exp\left\{ -\frac{tn^{\tau-1}(n-2u)}{4\alpha} \left[1 - \left(\frac{4(\sigma_{\xi,\mathbf{m}}^2 + \sigma_\varepsilon^2)}{\alpha} + \frac{64\sigma_{\xi,\mathbf{m}}^2 \sigma_\varepsilon^2}{\alpha^3} \right) \right] - \frac{\sigma_\varepsilon^2 u}{2\alpha} \left(1 - \frac{8\sigma_\varepsilon^2}{\alpha} \right) \right\}$$

since $m \in \mathcal{A}_t^{\mathsf{c}}$. Therefore,

$$\sum_{m \in \mathcal{A}_t} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)} \boldsymbol{\xi}_{\text{new},\mathbf{m}}^2 \right) \to 0.$$

Combining these calculations, it follows that

$$\limsup_{n\to\infty} n^{1-\tau} \sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)} \boldsymbol{\xi}_{\text{new},\mathbf{m}}^2 \right) \leq t.$$

Since $t > 0$ is arbitrary, this implies that

$$\sum_{\mathbf{m}\in\mathcal{M}} \mathbb{E}\left(\tilde{w}_{\mathbf{m}}^{(1)} \boldsymbol{\xi}_{\text{new},\mathbf{m}}^2 \right) = o(n^{\tau-1}),$$

which finishes the proof. $\square$

Proof of Theorem 3.4. Note that $(\mathbf{Z} \ominus \mathbf{W}) \oplus (\mathbf{Z}, \mathbf{W})^\perp \oplus \mathcal{C}(\mathbf{W}) = \mathbb{R}^n$, where $\mathcal{C}(\mathbf{W})$ denotes the column space of $\mathbf{W}$. Then, let $\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}} \in \mathbb{R}^{n\times r_{\mathbf{Z}\ominus\mathbf{W}}}$ (respectively $\mathbf{U}_{(\mathbf{Z},\mathbf{W})^\perp} \in \mathbb{R}^{n\times r_{(\mathbf{Z},\mathbf{W})^\perp}}$) be a matrix with orthonormal rows that spans $\mathbf{Z} \ominus \mathbf{W}$ (respectively $(\mathbf{Z}, \mathbf{W})^\perp$). Denote by $\mathbf{U} \in \mathbb{R}^{n\times(r_{\mathbf{Z}\ominus\mathbf{W}}+r_{(\mathbf{Z},\mathbf{W})^\perp})}$ the orthogonal matrix where $\mathbf{U} = (\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}, \mathbf{U}_{(\mathbf{Z},\mathbf{W})^\perp})$. By properties of the Gaussian distribution, $\mathbf{U}^\mathsf{T}\mathbf{X}$ has rows that are independent and identically distributed $\mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$ and the entries of $\mathbf{U}^\mathsf{T}\boldsymbol{\varepsilon}$ are independent. Thus, we partition $\mathbf{U}^\mathsf{T}\mathbf{X}$ into two parts, $\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}^\mathsf{T}\mathbf{X}$ and $\mathbf{U}_{(\mathbf{Z},\mathbf{W})^\perp}^\mathsf{T}\mathbf{X}$, with $\mathbf{U}_{(\mathbf{Z},\mathbf{W})^\perp}^\mathsf{T}\mathbf{X}$ being further decomposed into $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$. Thus, $\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}^\mathsf{T}\mathbf{X}$, $\tilde{\mathbf{X}}^{(1)}$, and $\tilde{\mathbf{X}}^{(2)}$ have independent and identically distributed rows. By Lemma A3.2, it follows that

$$\mathbb{E}\|\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}^\mathsf{T}\mathbf{X}(\tilde{\boldsymbol{\beta}}_{\text{EW}} - \boldsymbol{\beta})\|_2^2 = o(n^{\tau-1} r_{\mathbf{Z}\ominus\mathbf{W}}).$$

But, since $\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{U}_{\mathbf{Z}\ominus\mathbf{W}}^\mathsf{T} = \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}$, we have that

$$\mathbb{E}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}(\tilde{\boldsymbol{\beta}}_{\text{EW}} - \boldsymbol{\beta})\|_2^2 = o(n^{\tau-1} r_{\mathbf{Z}\ominus\mathbf{W}}).$$

Next, since $\boldsymbol{\nu}$ and $\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\boldsymbol{\varepsilon}$ are independent of $\tilde{\boldsymbol{\beta}}_{\text{EW}}$, standard arguments show that

$$(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}(\tilde{\boldsymbol{\beta}}_{\text{EW}}-\boldsymbol{\beta}) = o_{\mathbb{P}}(n^{\tau-1}r_{\mathbf{Z}\ominus\mathbf{W}}).$$

Therefore,

$$\begin{aligned}\|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}_{\text{EW}})\|_2^2 &= \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}(\tilde{\boldsymbol{\beta}}_{\text{EW}}-\boldsymbol{\beta})\|_2^2 + 2(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{X}(\tilde{\boldsymbol{\beta}}_{\text{EW}}-\beta) + \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 \\ &= \|\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}(\mathbf{Z}\boldsymbol{\nu}+\boldsymbol{\varepsilon})\|_2^2 + o_{\mathbb{P}}(n^{\tau-1}r_{\mathbf{Z}\ominus\mathbf{W}}).\end{aligned}$$

From Lemma A3.1, we have

$$\|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}(\mathbf{y}-\mathbf{X}\hat{\beta}_{\text{EW}})\|_2^2 = \|\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(n^{\tau}).$$

Now, the remainder of the proof is identical to that of Theorem 3.1 and is omitted, which finishes the proof. □

A.3.1.2 Proofs for Section ??

Lemma A3.3. Assume (3.4) and (3.5). Then, $\sigma_{\varsigma}^2 \asymp n$.

Proof. By a direct calculation, we have that

$$\begin{aligned}\sigma_{\varsigma}^2 = \text{Var}\left(\boldsymbol{\xi}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\xi}\right) &= \kappa_{\varepsilon}\sum_{i=1}^{n}\mathbf{Q}_{i,i}^2 + \kappa_{\nu}\sum_{i=n+1}^{n+q}\mathbf{Q}_{i,i}^2 \\ &\quad + 2\sum_{i\neq j}\mathbf{Q}_{i,j}^2(\sigma_{\varepsilon}^2\mathbb{1}_{1\le i\le n}+\sigma_{\nu}^2\mathbb{1}_{n+1\le i\le n+q}) \\ &\quad\quad \times(\sigma_{\varepsilon}^2\mathbb{1}_{1\le j\le n}+\sigma_{\nu}^2\mathbb{1}_{n+1\le j\le n+q}).\end{aligned}$$

Therefore,

$$\min(\kappa_{\varepsilon},\kappa_{\nu},2\sigma_{\varepsilon}^4,2\sigma_{\nu}^4)\|\mathbf{Q}\|_{\text{HS}}^2 \le \text{Var}\left(\boldsymbol{\xi}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\xi}\right) \le \max(\kappa_{\varepsilon},\kappa_{\nu},2\sigma_{\varepsilon}^4,2\sigma_{\nu}^4)\|\mathbf{Q}\|_{\text{HS}}^2.$$

Expanding $\mathbf{Q}^2$, we see that

$$\mathbf{Q}^2 = \begin{pmatrix}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} + r_{\mathbf{Z}\ominus\mathbf{W}}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}^{-2}\mathbf{P}^{\perp}_{(\mathbf{Z},\mathbf{W})} + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} & \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z} + \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} + \mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}} & \mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z} + \mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\end{pmatrix}$$

Thus,

$$\|\mathbf{Q}\|_{\text{HS}}^2 = \operatorname{tr}(\mathbf{Q}^2) = r_{\mathbf{Z}\ominus\mathbf{W}} + r_{\mathbf{Z}\ominus\mathbf{W}}^2 r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} + 2\operatorname{tr}(\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}) + \operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}).$$

Now,

$$r_{\mathbf{Z}\ominus\mathbf{W}} + r_{\mathbf{Z}\ominus\mathbf{W}}^2 r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} \le \|\mathbf{Q}\|_{\text{HS}}^2 \le r_{\mathbf{Z}\ominus\mathbf{W}} + r_{\mathbf{Z}\ominus\mathbf{W}}^2 r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1} + r_{\mathbf{Z}\ominus\mathbf{W}}\,\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T})(2+\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\mathsf{T})).$$

Invoking the two assumptions finishes the proof. □

Proof of Theorem 3.5. Recall from the calculations in Section 3.3.1 that

$$\hat{\sigma}_\nu^2 = [\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\left(\boldsymbol{\xi}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi} + o_{\mathbb{P}}(n^{1/2})\right).$$

Thus,

$$\sigma_\varsigma^{-1}\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})(\hat{\sigma}_\nu^2 - \sigma_\nu^2) = \sigma_\varsigma^{-1}(\boldsymbol{\xi}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi} - \sigma_\nu^2\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})) + \sigma_\varsigma^{-1}o_{\mathbb{P}}(n^{1/2}).$$

For the first term, noting that

$$\begin{aligned}
&\mathbb{E}\boldsymbol{\xi}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi} = \sigma_\nu^2\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z}),\\
&\operatorname{Var}(\boldsymbol{\xi}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi}) = \sigma_\varsigma^2,
\end{aligned}$$

we may apply Theorem 5.1 of Jiang (1996) to conclude that

$$\sigma_\varsigma^{-1}(\boldsymbol{\xi}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi} - \sigma_\nu^2[\operatorname{tr}(\mathbf{Z}^\mathsf{T}\mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

For the other term, we may invoke Lemma A3.3 to obtain

$$\sigma_\varsigma^{-1}o_{\mathbb{P}}(n^{1/2}) = o_{\mathbb{P}}(1).$$

This finishes the proof. □

Lemma A3.4. Consider the model given in equation (3.1.1). Assume (3.1), (3.2), and (3.5). Then,

$$\hat{\sigma}_\varepsilon^2 \xrightarrow{\mathbb{P}} \sigma_\varepsilon^2.$$

Proof of Lemma A3.4. Indeed, from Lemma A3.1,

$$\hat{\sigma}_\varepsilon^2 = r_{(\mathbf{Z},\mathbf{W})^\perp}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^\perp\boldsymbol{\varepsilon}\|_2^2 + o_{\mathbb{P}}(1).$$

Now, by the Hanson-Wright inequality (Theorem 1.1 of Rudelson and Vershynin (2013)), for any constant $a > 0$,

$$\mathbb{P}\left(|\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2^2 - \sigma_{\varepsilon}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}| > a\right) \leq 2\exp\left(-c\min\left(\frac{a^2}{K_{\varepsilon}^4 r_{(\mathbf{Z},\mathbf{W})^{\perp}}}, \frac{a}{K_{\varepsilon}^2}\right)\right) \qquad \text{(A.3.1.2)}$$

where $c > 0$ is a universal constant. Setting $a = r_{(\mathbf{Z},\mathbf{W})^{\perp}}^{3/4}$, it follows that

$$\mathbb{P}\left(|\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2^2 - \sigma_{\varepsilon}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}| > r_{(\mathbf{Z},\mathbf{W})^{\perp}}^{3/4}\right) \to 0.$$

This implies that

$$r_{(\mathbf{Z},\mathbf{W})^{\perp}}^{-1}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2^2 \xrightarrow{\mathbb{P}} \sigma_{\varepsilon}^2,$$

which finishes the proof. □

Proof of Proposition 3.6. Temporarily, let $\Delta = \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}(\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_4$. Then,

$$\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_4 - \Delta \leq \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_4 \leq \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_4 + \Delta. \qquad \text{(A.3.1.3)}$$

Now, applying Lemma 6.3 of Law and Ritov (2021b),

$$\Delta \leq \|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}(\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_2 \leq \|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}}\|_2 = o_{\mathbb{P}}(n^{1/4}).$$

Let t_n be a sequence depending on n that will be chosen later. Define the event $\mathscr{T}$ as

$$\mathscr{T} \triangleq \{\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_4 \geq t_n n^{1/4}\}.$$

By norm equivalence, we have that

$$\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_4 \geq n^{-1/4}\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2.$$

Setting $a = \sigma_{\varepsilon}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}/2$ in equation (A.3.1.2), it follows that

$$\mathbb{P}\left(\|\mathbf{P}_{(\mathbf{Z},\mathbf{W})}^{\perp}\boldsymbol{\varepsilon}\|_2^2 \geq \sigma_{\varepsilon}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}/2\right) \leq 2\exp\left(-c\min\left(\frac{\sigma_{\varepsilon}^2 r_{(\mathbf{Z},\mathbf{W})^{\perp}}}{4K_{\varepsilon}^4}, \frac{r_{(\mathbf{Z},\mathbf{W})^{\perp}}}{2K_{\varepsilon}^2}\right)\right).$$

Therefore, choosing $t_n = \sqrt{(\sigma_\varepsilon^2 r_{(\mathbf{Z},\mathbf{W})^\perp})/(2n)}$, we have that

$$\mathbb{P}(\mathscr{T}) \geq 1 - 2\exp\left(-c\min\left(\frac{\sigma_\varepsilon^2 r_{(\mathbf{Z},\mathbf{W})^\perp}}{4K_\nu^4}, \frac{r_{(\mathbf{Z},\mathbf{W})^\perp}}{2K_\nu^2}\right)\right) \to 1.$$

Note that $\liminf_{n\to\infty} t_n > 0$ by assumption (3.5). Thus, on $\mathscr{T}$ for n sufficiently large, it follows that

$$\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4 - \Delta \geq 0.$$

Now, raising all terms in equation (A.3.1.3) to the fourth power, we see that

$$\left(\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4 - \Delta\right)^4 \mathbb{1}_{\mathscr{T}} \leq \|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_4^4 \mathbb{1}_{\mathscr{T}} \leq \left(\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4 + \Delta\right)^4 \mathbb{1}_{\mathscr{T}}.$$

Expanding the left and right hand side and using the fact that $\Delta = o_{\mathbb{P}}(n^{1/4})$, the above implies that

$$\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{EW}})\|_4^4 \mathbb{1}_{\mathscr{T}} = \|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4^4 \mathbb{1}_{\mathscr{T}} + o_{\mathbb{P}}(n).$$

Recalling that $n = mq$, Lemma A3.4 and the above calculations show that

$$\hat{\omega}_\varepsilon \mathbb{1}_{\mathscr{T}} = q^{-1}m^2\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4^4 \mathbb{1}_{\mathscr{T}} - 3(m-1)\sigma_\varepsilon^4 \mathbb{1}_{\mathscr{T}} + o_{\mathbb{P}}(1).$$

A direct calculation yields

$$q^{-1}m^2\mathbb{E}\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\varepsilon\|_4^4 = 3(m-1)\sigma_\varepsilon^4 + \omega_\varepsilon.$$

Now, since $\mathbf{Z} = \mathbf{I}_q \otimes \mathbf{1}_m$, $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}$ is a block diagonal matrix, with q blocks of $m^{-1}\mathbf{1}_m\mathbf{1}_m^\mathsf{T}$. Hence, we may partition $\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}$ into q blocks of length m, whereby each block is independent and identically distributed. Then, it follows by a law of large numbers that

$$q^{-1}m^2\|\mathbf{P}^\perp_{(\mathbf{Z},\mathbf{W})}\boldsymbol{\varepsilon}\|_4^4 \xrightarrow{\mathbb{P}} 3(m-1)\sigma_\varepsilon^4 + \omega_\varepsilon.$$

Thus,

$$(\hat{\omega}_\varepsilon - \omega_\varepsilon)\mathbb{1}_{\mathscr{T}} = o_{\mathbb{P}}(1).$$

Since $\mathbb{P}(\mathscr{T}) \to 1$, this proves the claim. $\square$

A.3.1.3 Proofs for Section ??

Proof of Proposition 3.7. Indeed, from the calculations in Section 3.3.1, we have that

$$\tilde{\sigma}_\nu^2 = [\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\left(\boldsymbol{\xi}^\top \mathbf{Q}\boldsymbol{\xi} + o_{\mathbb{P}}(n)\right).$$

By a variance calculation, we have that

$$\mathrm{Var}([\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\boldsymbol{\xi}^\top \mathbf{Q}\boldsymbol{\xi}) = \sigma_\varsigma^2[\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-2}.$$

Applying Lemma A3.3 and assumption (3.10) shows that $\mathrm{Var}([\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\boldsymbol{\xi}^\top \mathbf{Q}\boldsymbol{\xi}) \to 0$. Noting that $[\mathrm{tr}(\mathbf{Z}^\top \mathbf{P}_{\mathbf{Z}\ominus\mathbf{W}}\mathbf{Z})]^{-1}\mathbb{E}\boldsymbol{\xi}^\top\boldsymbol{\xi} = \sigma_\nu^2$ proves the first claim. The proof for $\tilde{\sigma}_\gamma^2$ is analogous. Finally, the last claim that $\tilde{\sigma}_\varepsilon^2 \xrightarrow{\mathbb{P}} \sigma_\varepsilon^2$ is identical to the proof of Lemma A3.4, which finishes the proof. □

Proof of Lemma 3.8. By independence, the joint distribution of $(\nu, \gamma, \varepsilon)$ is given by

$$\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\gamma}\\ \boldsymbol{\varepsilon}\end{pmatrix} \sim \mathcal{N}_{v+r+n}\left(\begin{pmatrix}\mathbf{0}_v\\ \mathbf{0}_r\\ \mathbf{0}_n\end{pmatrix}, \begin{pmatrix}\sigma_\nu^2\mathbf{I}_v & \mathbf{0}_{v\times r} & \mathbf{0}_{v\times n}\\ \mathbf{0}_{r\times v} & \sigma_\gamma^2\mathbf{I}_r & \mathbf{0}_{r\times n}\\ \mathbf{0}_{n\times v} & \mathbf{0}_{n\times r} & \sigma_\varepsilon^2\mathbf{I}_n\end{pmatrix}\right).$$

Therefore, the joint distribution of Y and η is given by

$$\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\eta}\end{pmatrix} \sim \mathcal{N}_{2n}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{pmatrix}, \begin{pmatrix}\sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top + \sigma_\gamma^2\mathbf{W}\mathbf{W}^\top + \sigma_\varepsilon^2\mathbf{I}_n & \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top\\ \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top & \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top\end{pmatrix}\right).$$

By standard results on the conditional mean of $\boldsymbol{\eta}$ given $\mathbf{y}$, it follows that

$$\mathbb{E}\left(\boldsymbol{\eta}|\mathbf{y}\right) = \boldsymbol{\mu} + \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top\left(\sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top + \sigma_\gamma^2\mathbf{W}\mathbf{W}^\top + \sigma_\varepsilon^2\mathbf{I}_n\right)^{-1}(\mathbf{y} - \boldsymbol{\mu}).$$

which finishes the proof. □

Proof of Theorem 3.9. From Proposition 3.7, it follows that

$$\tilde{\sigma}_\nu^2 = \sigma_\nu^2 + \delta_\nu^2 \qquad \tilde{\sigma}_\gamma^2 = \sigma_\gamma^2 + \delta_\gamma^2 \qquad \tilde{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + \delta_\varepsilon^2,$$

where δ_ν^2, δ_γ^2, and δ_ε^2 are all $o_{\mathbb{P}}(1)$. We write $\delta_*^2 \triangleq \max\{\delta_\nu^2, \delta_\gamma^2, \delta_\varepsilon^2\}$. Define

$$\begin{aligned}\boldsymbol{\Sigma} &\triangleq \sigma_\nu^2\mathbf{Z}\mathbf{Z}^\top + \sigma_\gamma^2\mathbf{W}\mathbf{W}^\top + \sigma_\varepsilon^2\mathbf{I}_n,\\ \boldsymbol{\Delta} &\triangleq \delta_\nu^2\mathbf{Z}\mathbf{Z}^\top + \delta_\gamma^2\mathbf{W}^\top + \delta_\varepsilon^2\mathbf{I}_n.\end{aligned}$$

Note that $\mathbf{\Sigma}$ is positive definite and $\mathbf{\Delta}$ is invertible almost surely. Then, by the Matrix Inversion Lemma,

$$(\mathbf{\Sigma}+\mathbf{\Delta})^{-1}=\mathbf{\Sigma}^{-1}-\mathbf{\Sigma}^{-1}\left(\mathbf{\Delta}^{-1}+\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{\Sigma}^{-1}.$$

Some algebra yields

$$\begin{aligned}
\tilde{\boldsymbol{\eta}}_{\text{EW}} =&\tilde{\boldsymbol{\mu}}_{\text{EW}}+\tilde{\sigma}_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}\left(\tilde{\sigma}_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}+\tilde{\sigma}_{\gamma}^{2}\mathbf{W}\mathbf{W}^{\intercal}+\tilde{\sigma}_{\varepsilon}^{2}\mathbf{I}_{n}\right)^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\\
=&\tilde{\boldsymbol{\mu}}_{\text{EW}}+\left(\sigma_{\nu}^{2}+\delta_{\nu}^{2}\right)\mathbf{Z}\mathbf{Z}^{\intercal}(\mathbf{\Sigma}+\mathbf{\Delta})^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\\
=&(\tilde{\boldsymbol{\mu}}_{\text{EW}}-\boldsymbol{\mu})+\left(\sigma_{\nu}^{2}+\delta_{\nu}^{2}\right)\mathbf{Z}\mathbf{Z}^{\intercal}(\mathbf{\Sigma}+\mathbf{\Delta})^{-1}(\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}}_{\text{EW}})\\
&+\delta_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}(\mathbf{\Sigma}+\mathbf{\Delta})^{-1}(\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon})\\
&-\sigma_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}\mathbf{\Sigma}^{-1}\left(\mathbf{\Delta}^{-1}+\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon})\\
&+\boldsymbol{\mu}+\sigma_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}),
\end{aligned}$$

where we have added and subtracted μ and applied the Matrix Inversion Lemma. From Lemma 3.8, recall that

$$\tilde{\boldsymbol{\eta}}_{\text{oracle}}=\boldsymbol{\mu}+\sigma_{\nu}^{2}\mathbf{Z}\mathbf{Z}^{\intercal}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}).$$

Define $\boldsymbol{\xi}\triangleq\tilde{\boldsymbol{\eta}}_{\text{EW}}-\tilde{\boldsymbol{\eta}}_{\text{oracle}}$. Therefore,

$$n^{-1}\left(\|\tilde{\boldsymbol{\eta}}_{\text{EW}}-\boldsymbol{\eta}\|_{2}^{2}-\|\tilde{\boldsymbol{\eta}}_{\text{oracle}}-\boldsymbol{\eta}\|_{2}^{2}\right)=n^{-1}\left(\|\boldsymbol{\xi}\|_{2}^{2}+2\boldsymbol{\xi}^{\intercal}(\tilde{\boldsymbol{\eta}}_{\text{oracle}}-\boldsymbol{\eta})\right).$$

We will prove each of the two terms on the right hand side are $o_{\mathbb{P}}(1)$. Before doing so, we prove a few useful facts to facilitate the remainder of the proof.

(I) $\|\mathbf{\Sigma}\|_{2}^{2}=\mathcal{O}(1)$.

(II) $\|\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}\|_{2}^{2}=\mathcal{O}_{\mathbb{P}}(n)$ and $\|\mathbf{Z}\boldsymbol{\nu}\|_{2}^{2}=\mathcal{O}_{\mathbb{P}}(n)$.

(III) $\|\mathbf{\Delta}\|_{2}^{2}=o_{\mathbb{P}}(1)$ and $\|(\mathbf{\Sigma}+\mathbf{\Delta})^{-1}\|_{2}^{2}=\mathcal{O}_{\mathbb{P}}(1)$.

(I) is immediate since $\mathbf{\Sigma}$ by the triangle inequality and assumptions (3.4) and (3.11). The first part of (II) follows from the fact that

$$\|\mathbf{Z}\boldsymbol{\nu}+\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}\|_{2}^{2}\preceq\lambda_{\max}(\mathbf{\Sigma})\,\chi_{n}^{2}=\mathcal{O}_{\mathbb{P}}(n),$$

where $\preceq$ denotes stochastic ordering and $\lambda_{\max}(\mathbf{\Sigma})=\mathcal{O}(1)$ by assumptions (3.4) and (3.11). The second part of (II) is similar. Finally, for (III), the first part follows from $\delta_{*}^{2}\xrightarrow{\mathbb{P}}0$ and $\|\mathbf{\Sigma}\|_{2}=\mathcal{O}(1)$.

For the second part, note that the minimal singular value of $\mathbf{\Sigma}$ is bounded away from zero. Since $\delta_*^2 \xrightarrow{\mathbb{P}} 0$, it follows that for n sufficiently large, the minimal singular value of $\mathbf{\Sigma} + \mathbf{\Delta}$ is bounded away from zero. This proves all three claims. We can now show that $\|\boldsymbol{\xi}\|_2^2 = o_{\mathbb{P}}(n)$, which we show in parts. To this end, note that

$$\begin{aligned}
&\|(\mathbf{I}_n - (\sigma_\nu^2 + \delta_\nu^2)\mathbf{Z}\mathbf{Z}^\mathsf{T}(\mathbf{\Sigma} + \mathbf{\Delta})^{-1})(\tilde{\boldsymbol{\mu}}_{\text{EW}} - \boldsymbol{\mu})\|_2^2 \\
&\leq \|\mathbf{I}_n - (\sigma_\nu^2 + \delta_\nu^2)\mathbf{Z}\mathbf{Z}^\mathsf{T}(\mathbf{\Sigma} + \mathbf{\Delta})^{-1}\|_2^2 \|\tilde{\boldsymbol{\mu}}_{\text{EW}} - \boldsymbol{\mu}\|_2^2 \\
&= o_{\mathbb{P}}(n).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|\delta_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T}(\mathbf{\Sigma} + \mathbf{\Delta})^{-1}(\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\|_2^2 \\
&\leq \delta_\nu^4 \|\mathbf{Z}\mathbf{Z}^\mathsf{T}\|_2^2 \|(\mathbf{\Sigma} + \mathbf{\Delta})^{-1}\|_2^2 \|\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\|_2^2 \\
&= o_{\mathbb{P}}(1).
\end{aligned}$$

For the last term, we apply the Matrix Inversion Lemma again to obtain

$$(\mathbf{\Delta}^{-1} + \mathbf{\Sigma}^{-1})^{-1} = \mathbf{\Delta} - \mathbf{\Delta}(\mathbf{\Delta} + \mathbf{\Sigma})^{-1}\mathbf{\Delta}.$$

Hence,

$$\begin{aligned}
&\|\sigma_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{\Delta}^{-1} + \mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\|_2^2 \\
&\leq \sigma_\nu^4 \|\mathbf{Z}\mathbf{Z}^\mathsf{T}\|_2^2 \|\mathbf{\Sigma}^{-1}\|_2^4 \|\mathbf{\Delta}\|_2^2 \|\mathbf{I}_n - (\mathbf{\Delta} + \mathbf{\Sigma})^{-1}\mathbf{\Delta}\|_2^2 \|\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\|_2^2 \\
&= o_{\mathbb{P}}(n).
\end{aligned}$$

Combining these three results with the triangle inequality, this proves that

$$\|\boldsymbol{\xi}\|_2^2 = o_{\mathbb{P}}(n).$$

For the other quantity, we have that

$$\begin{aligned}
\|\tilde{\boldsymbol{\eta}}_{\text{oracle}} - \boldsymbol{\eta}\|_2^2 &= \|\sigma_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) - \mathbf{Z}\boldsymbol{\nu}\|_2^2 \\
&\leq 2\|\sigma_\nu^2 \mathbf{Z}\mathbf{Z}^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\|_2^2 + 2\|\mathbf{Z}\boldsymbol{\nu}\|_2^2 \\
&\leq 2\sigma_\nu^4 \|\mathbf{Z}\mathbf{Z}^\mathsf{T}\|_2^2 \|\mathbf{\Sigma}^{-1}\|_2^2 \|\mathbf{Z}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\|_2^2 + 2\|\mathbf{Z}\boldsymbol{\nu}\|2^2.
\end{aligned}$$

Applying all three facts from above demonstrates that

$$\|\tilde{\boldsymbol{\eta}}_{\text{oracle}} - \boldsymbol{\eta}\|_2^2 = \mathcal{O}_{\mathbb{P}}(n).$$

Using the Cauchy-Schwarz inequality yields

$$2|\xi^{\mathsf{T}}(\tilde{\boldsymbol{\eta}}_{\text{oracle}} - \boldsymbol{\eta})| \leq 2\|\boldsymbol{\xi}\|_2\|\tilde{\boldsymbol{\eta}}_{\text{oracle}} - \boldsymbol{\eta}\|_2 = o_{\mathbb{P}}(n),$$

which finishes the proof. □

A.3.2 Additional Simulation Results

In this section, we include all the simulation tables for Section 5.4.

Table A.3.1: Simulations with $d = 0$ and $s = 3$

Simulations with $d = 0$ and $s = 3$

	σ_ν^2	0	0	0	0	1	1	1	1
	Distr	z	z	e	e	z	z	e	e
	ρ	0.0	0.8	0.0	0.8	0.0	0.8	0.0	0.8
Type I/II Error	EW_G	0.04	0.03	0.05	0.04	0.00	0.00	0.00	0.00
	SL_G	0.04	0.03	0.04	0.04	0.00	0.00	0.00	0.00
	LS_G	0.03	0.03	0.05	0.05	0.00	0.00	0.00	0.00
	LD	0.02	0.01	0.02	0.02	0.00	0.00	0.00	0.00
	EW_SG	0.12	0.12	0.09	0.11	0.00	0.00	0.00	0.00
	SL_SG	0.13	0.11	0.09	0.12	0.00	0.00	0.00	0.00
	LS_SG	0.10	0.10	0.09	0.10	0.00	0.00	0.00	0.00
Average Coverage	EW_G	1.00	1.00	0.99	1.00	0.93	0.93	0.81	0.81
	SL_G	1.00	1.00	0.99	1.00	0.94	0.92	0.82	0.82
	LS_G	1.00	1.00	0.99	1.00	0.94	0.93	0.82	0.82
	LD	0.98	0.99	0.97	0.98	0.97	0.96	0.87	0.86
	EW_SG	0.95	0.96	0.95	0.96	0.89	0.91	0.89	0.90
	SL_SG	0.95	0.95	0.95	0.94	0.90	0.90	0.89	0.89
	LS_SG	0.94	0.95	0.95	0.95	0.92	0.91	0.90	0.91
Average Length	EW_G	0.04	0.04	0.04	0.04	0.43	0.43	0.43	0.42
	SL_G	0.04	0.04	0.04	0.05	0.43	0.43	0.43	0.43
	LS_G	0.04	0.04	0.04	0.04	0.43	0.43	0.43	0.43
	LD	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24
	EW_SG	0.04	0.04	0.04	0.04	0.42	0.42	0.60	0.60
	SL_SG	0.04	0.04	0.04	0.04	0.42	0.42	0.61	0.60
	LS_SG	0.04	0.04	0.04	0.04	0.42	0.42	0.61	0.61
Average Loss	EW	0.01	0.02	0.01	0.02	0.19	0.19	0.18	0.19
	SL	0.03	0.03	0.03	0.02	0.21	0.20	0.21	0.19
	LS	0.01	0.01	0.01	0.01	0.17	0.17	0.17	0.17
	LD	0.00	0.00	0.00	0.00	0.17	0.17	0.17	0.17

Table A.3.2: Simulations with $d = 200$ and $s = 3$

Simulations with $d = 200$ and $s = 3$

	σ_ν^2	0	0	0	0	1	1	1	1
	Distr	z	z	e	e	z	z	e	e
	ρ	0.0	0.8	0.0	0.8	0.0	0.8	0.0	0.8
Type I/II Error	$\mathrm{EW_G}$	0.04	0.03	0.05	0.04	0.00	0.00	0.00	0.00
	$\mathrm{SL_G}$	0.04	0.03	0.04	0.04	0.00	0.00	0.00	0.00
	$\mathrm{LS_G}$	0.03	0.03	0.05	0.05	0.00	0.00	0.00	0.00
	LD	0.02	0.01	0.02	0.02	0.00	0.00	0.00	0.00
Average Coverage	$\mathrm{EW_G}$	1.00	1.00	0.99	1.00	0.93	0.93	0.81	0.81
	$\mathrm{SL_G}$	1.00	1.00	0.99	1.00	0.94	0.92	0.82	0.82
	$\mathrm{LS_G}$	1.00	1.00	0.99	1.00	0.94	0.93	0.82	0.82
	LD	0.98	0.99	0.97	0.98	0.97	0.96	0.87	0.86
Average Length	$\mathrm{EW_G}$	0.04	0.04	0.04	0.04	0.43	0.43	0.43	0.42
	$\mathrm{SL_G}$	0.04	0.04	0.04	0.05	0.43	0.43	0.43	0.43
	$\mathrm{LS_G}$	0.04	0.04	0.04	0.04	0.43	0.43	0.43	0.43
	LD	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24
Average Loss	EW	0.01	0.02	0.01	0.02	0.19	0.19	0.18	0.19
	SL	0.03	0.03	0.03	0.02	0.21	0.20	0.21	0.19
	LS	0.01	0.01	0.01	0.01	0.17	0.17	0.17	0.17
	LD	0.00	0.00	0.00	0.00	0.17	0.17	0.17	0.17

Table A.3.3: Simulations with Gaussian errors when $s = 15$

Simulations with Gaussian errors when $s = 15$

	σ_ν^2	0	0	0	0	1	1	1	1
	d	0	0	200	200	0	0	200	200
	ρ	0	0.8	0	0.8	0	0.8	0	0.8
Type I/II Error	EW_G	0.05	0.05	0.05	0.05	0.00	0.00	0.00	0.00
	SL_G	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00
	LS_G	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00
	LD	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.00
	EW_SG	0.12	0.09	-	-	0.00	0.00	-	-
	SL_SG	0.15	0.11	-	-	0.00	0.00	-	-
	LS_SG	0.12	0.09	-	-	0.00	0.00	-	-
Average Coverage	EW_G	0.99	0.99	0.99	0.99	0.85	0.84	0.85	0.84
	SL_G	0.98	0.99	0.98	0.99	0.86	0.88	0.86	0.88
	LS_G	0.99	0.99	0.99	0.99	0.89	0.90	0.89	0.90
	LD	0.97	0.97	0.97	0.97	0.95	0.95	0.95	0.95
	EW_SG	0.94	0.95	-	-	0.84	0.82	-	-
	SL_SG	0.95	0.95	-	-	0.86	0.89	-	-
	LS_SG	0.95	0.96	-	-	0.90	0.91	-	-
Average Length	EW_G	0.05	0.05	0.05	0.05	0.42	0.42	0.42	0.42
	SL_G	0.05	0.06	0.05	0.06	0.44	0.45	0.44	0.45
	LS_G	0.05	0.05	0.05	0.05	0.43	0.43	0.43	0.43
	LD	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24
	EW_SG	0.04	0.04	-	-	0.42	0.41	-	-
	SL_SG	0.05	0.06	-	-	0.43	0.45	-	-
	LS_SG	0.04	0.04	-	-	0.42	0.42	-	-
Average Loss	EW	0.04	0.06	0.04	0.06	0.22	0.25	0.22	0.25
	SL	0.17	0.45	0.17	0.45	0.43	0.61	0.43	0.61
	LS	0.02	0.02	0.02	0.02	0.19	0.19	0.19	0.19
	LD	0.00	0.00	0.00	0.00	0.17	0.17	0.17	0.17

APPENDIX 3

Appendix of Chapter 4

In Section 4.3, we use the following assumptions.

A.4.1 Some Technical Lemmata

In this section, we provide some technical lemmata regarding the trigonometric basis that are used in the proofs. We start with a lemma regarding the fourth moments of the trigonometric basis functions.

Lemma A4.1. Let $a, b, k, l \in \mathbb{N}$ be fixed positive integers and $(\varphi_k)_{k=1}^{\infty}$ denote the trigonometric basis as defined in Definition 4.1.1. Then,

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad \leq \delta_{a+b,k+l} + \delta_{a+b+k,l} + \delta_{a+b+l,k} + \delta_{a+k+l,b} + \delta_{a,b+k+l} + \delta_{a+k,b+l} + \delta_{a+l,b+k} \\
&\qquad\quad + \delta_{l,1}\left(\delta_{a+b,k} + \delta_{a+k,b} + \delta_{a,b+k}\right) + \delta_{k,1}\left(\delta_{a+b,l} + \delta_{a+l,b} + \delta_{a,b+l}\right) \\
&\qquad\quad + \delta_{b,1}\left(\delta_{a+k,l} + \delta_{a+l,k} + \delta_{a,k+l}\right) + \delta_{a,1}\left(\delta_{b+k,l} + \delta_{b+l,k} + \delta_{b,k+l}\right).
\end{aligned}$$

Proof. We consider a few cases:

1. $a, b, k, l > 1$ and a, b, k, l are even.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 4\int_0^1 \cos(\pi at)\cos(\pi bt)\cos(\pi kt)\cos(\pi lt)dt \\
&\qquad = \int_0^1 (\cos(\pi(a+b)t) + \cos(\pi(a-b)t)) \\
&\qquad\qquad \times (\cos(\pi(k+l)t) + \cos(\pi(k-l)t))dt \\
&\qquad =\frac{1}{2}\int_0^1 (\cos(\pi(a+b+k+l)t) + \cos(\pi(a+b-k-l)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\cos(\pi(a+b+k-l)t) + \cos(\pi(a+b-k+l)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\cos(\pi(a-b+k+l)t) + \cos(\pi(a-b-k-l)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\cos(\pi(a-b+k-l)t) + \cos(\pi(a-b-k+l)))\, dt \\
&\qquad = \frac{1}{2}(\delta_{a+b,k+l} + \delta_{a+b+k,l} + \delta_{a+b+l,k} \\
&\qquad\qquad + \delta_{a+k+l,b} + \delta_{a,b+k+l} + \delta_{a+k,b+l} + \delta_{a+l,b+k}).
\end{aligned}$$

2. $a, b, k, l > 1$, a, b, k are even, and l is odd.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 4\int_0^1 \cos(\pi at)\cos(\pi bt)\cos(\pi kt)\sin(\pi(l-1)t)dt \\
&\qquad = \int_0^1 (\cos(\pi(a+b)t) + \cos(\pi(a-b)t)) \\
&\qquad\qquad \times (\sin(\pi(k+l-1)t) - \sin(\pi(k-l+1)t))dt \\
&\qquad =\frac{1}{2}\int_0^1 (\sin(\pi(a+b+k+l-1)t) + \sin(\pi(a+b-k-l+1)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\sin(\pi(a+b+k-l+1)t) + \sin(\pi(a+b-k+l-1)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\sin(\pi(a-b+k+l-1)t) + \sin(\pi(a-b-k-l+1)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\sin(\pi(a-b+k-l+1)t) + \sin(\pi(a-b-k+l-1)))\, dt \\
&\qquad = 0.
\end{aligned}$$

3. $a, b, k, l > 1$, a, b are even, and k, l are odd.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 4\int_0^1 \cos(\pi a t)\cos(\pi b t)\sin(\pi(k-1)t)\sin(\pi(l-1)t)dt \\
&\qquad = \int_0^1 (\cos(\pi(a+b)t) + \cos(\pi(a-b)t)) \\
&\qquad\qquad \times (\cos(\pi(k-l)t) - \cos(\pi(k+l-2)t))dt \\
&\qquad =\frac{1}{2}\int_0^1 (\cos(\pi(a+b+k-l)t) + \cos(\pi(a+b-k+l)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\cos(\pi(a+b+k+l-2)t) + \cos(\pi(a+b-k-l+2)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\cos(\pi(a-b+k-l)t) + \cos(\pi(a-b-k+l)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\cos(\pi(a-b+k+l-2)t) + \cos(\pi(a-b-k-l+2)))\, dt \\
&\qquad \leq \frac{1}{2}(\delta_{a+b+k,l} + \delta_{a+b+l,k} + \delta_{a+k,b+l} + \delta_{a+l,b+k}).
\end{aligned}$$

4. $a, b, k, l > 1$, a is even, and b, k, l are odd.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 4\int_0^1 \cos(\pi a t)\sin(\pi(b-1)t)\sin(\pi(k-1)t)\sin(\pi(l-1)t)dt \\
&\qquad = \int_0^1 (\sin(\pi(a+b-1)t) - \sin(\pi(a-b+1)t)) \\
&\qquad\qquad \times (\cos(\pi(k-l)t) - \cos(\pi(k+l-2)t))dt \\
&\qquad =\frac{1}{2}\int_0^1 (\sin(\pi(a+b+k-l-1)t) + \sin(\pi(a+b-k+l-1)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\sin(\pi(a+b+k+l-3)t) + \sin(\pi(a+b-k-l+1)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\sin(\pi(a-b+k-l+1)t) + \sin(\pi(a-b-k+l+1)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\sin(\pi(a-b+k+l-1)t) + \sin(\pi(a-b-k-l+3)))\, dt \\
&\qquad = 0.
\end{aligned}$$

5. $a, b, k, l > 1$ and a, b, k, l are odd.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 4\int_0^1 \sin(\pi(a-1)t)\sin(\pi(b-1)t)\sin(\pi(k-1)t)\sin(\pi(l-1)t)dt \\
&\qquad = \int_0^1 (\cos(\pi(a-b)t) - \cos(\pi(a+b-2)t)) \\
&\qquad\qquad \times (\cos(\pi(k-l)t) - \cos(\pi(k+l-2)t))dt \\
&\qquad =\frac{1}{2}\int_0^1 (\cos(\pi(a-b+k-l)t) + \cos(\pi(a-b-k+l)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\cos(\pi(a-b+k+l-2)t) + \cos(\pi(a-b-k-l+2)))\, dt \\
&\qquad\quad - \frac{1}{2}\int_0^1 (\cos(\pi(a+b+k-l-2)t) + \cos(\pi(a+b-k+l-2)))\, dt \\
&\qquad\quad + \frac{1}{2}\int_0^1 (\cos(\pi(a+b+k+l-4)t) + \cos(\pi(a+b-k-l)))\, dt \\
&\qquad \le \frac{1}{2}(\delta_{a+b,k+l} + \delta_{a+k,b+l} + \delta_{a+l,b+k}).
\end{aligned}$$

6. $a, b, k > 1$, a, b, k are even, and $l = 1$.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 2\sqrt{2}\int_0^1 \cos(\pi at)\cos(\pi bt)\cos(\pi kt)dt \\
&\qquad = \sqrt{2}\int_0^1 (\cos(\pi(a+b)t) + \cos(\pi(a-b)t))\cos(\pi kt)dt \\
&\qquad =\frac{1}{\sqrt{2}}\int_0^1 (\cos(\pi(a+b+k)t) + \cos(\pi(a+b-k)))\, dt \\
&\qquad\quad + \frac{1}{\sqrt{2}}\int_0^1 (\cos(\pi(a-b+k)t) + \cos(\pi(a-b-k)))\, dt \\
&\qquad = \frac{1}{\sqrt{2}}(\delta_{a+b,k} + \delta_{a+k,b} + \delta_{a,b+k}).
\end{aligned}$$

7. $a, b, k > 1$, a, b are even, k is odd, and $l = 1$.

$$
\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 2\sqrt{2}\int_0^1 \cos(\pi a t)\cos(\pi b t)\sin(\pi(k-1)t)dt \\
&\qquad = \sqrt{2}\int_0^1 (\cos(\pi(a+b)t)+\cos(\pi(a-b)t))\sin(\pi(k-1)t)dt \\
&\qquad =\frac{1}{\sqrt{2}}\int_0^1 \left(\sin(\pi(a+b+k-1)t)-\sin(\pi(a+b-k+1))\right)dt \\
&\qquad\quad + \frac{1}{\sqrt{2}}\int_0^1 \left(\sin(\pi(a-b+k-1)t)-\sin(\pi(a-b-k+1))\right)dt \\
&\qquad = 0.
\end{aligned}
$$

8. $a, b, k > 1$, a is even, b, k are odd, and $l = 1$.

$$
\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 2\sqrt{2}\int_0^1 \cos(\pi a t)\sin(\pi(b-1)t)\sin(\pi(k-1)t)dt \\
&\qquad = \sqrt{2}\int_0^1 \cos(\pi a t)(\cos(\pi(b-k)t)-\cos(\pi(b+k-2)t))dt \\
&\qquad =\frac{1}{\sqrt{2}}\int_0^1 \left(\cos(\pi(a+b-k)t)+\cos(\pi(a-b+k))\right)dt \\
&\qquad\quad - \frac{1}{\sqrt{2}}\int_0^1 \left(\cos(\pi(a+b+k-2)t)+\cos(\pi(a-b-k+2))\right)dt \\
&\qquad \leq \frac{1}{\sqrt{2}}(\delta_{a+b,k}+\delta_{a+k,b}).
\end{aligned}
$$

9. $a, b, k > 1$, a, b, k are odd, and $l = 1$.

$$\begin{aligned}
&\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt \\
&\qquad = 2\sqrt{2}\int_0^1 \sin(\pi(a-1)t)\sin(\pi(b-1)t)\sin(\pi(k-1)t)dt \\
&\qquad = \sqrt{2}\int_0^1 \sin(\pi(a-1)t)(\cos(\pi(b-k)t) - \cos(\pi(b+k-2)t))dt \\
&\qquad = \frac{1}{\sqrt{2}}\int_0^1 \left(\sin(\pi(a+b-k-1)t) + \sin(\pi(a-b+k-1))\right)dt \\
&\qquad\quad - \frac{1}{\sqrt{2}}\int_0^1 \left(\sin(\pi(a+b+k-3)t) + \sin(\pi(a-b-k+1))\right)dt \\
&\qquad = 0.
\end{aligned}$$

10. $a, b > 1$ and $k, l = 1$. $\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt = \delta_{a,b}$.

11. $a > 1$ and $b, k, l = 1$. $\int_0^1 \varphi_a(t)\varphi_b(t)\varphi_k(t)\varphi_l(t)dt = \delta_{a,1}$.

Combining these cases together and considering all permutations finishes the proof. $\square$

Next, we have a lemma regarding the aliasing effect in Fourier transforms on the uniform grid.

Lemma A4.2. For k even and $m \in \mathbb{N}$,

$$m^{-1}\sum_{j=1}^{m}\exp\left(\mathrm{i}\pi kj/m\right) = \begin{cases} 1, & \text{if } k = cm \text{ for } c \text{ even}, \\ 0, & \text{else.} \end{cases}$$

Proof of Lemma A4.2. Suppose that $k = cm$ for $c \in \mathbb{Z}$. Then, we have that $\exp(\mathrm{i}\pi kj/m) = \exp(\mathrm{i}\pi cj) = (-1)^{cj}$ for all $j = 1, \dots, m$. If c is even, then

$$m^{-1}\sum_{j=1}^{m}\exp\left(\mathrm{i}\pi kj/m\right) = m^{-1}\sum_{j=1}^{m} 1 = 1.$$

Now, if c is odd, this implies that m is even since k is even. Thus,

$$m^{-1}\sum_{j=1}^{m}\exp\left(\mathrm{i}\pi kj/m\right) = m^{-1}\sum_{j=1}^{m}(-1)^j = 0.$$

Finally, suppose that $k \neq cm$ for any $c \in \mathbb{Z}$. In this setting, we have that $\exp(\mathrm{i}\pi k/m) \neq 1$ while

$\exp(\mathfrak{i}\pi k) = 1$. Therefore,

$$m^{-1}\sum_{j=1}^{m}\exp(\mathfrak{i}\pi kj/m) = m^{-1}\exp(\mathfrak{i}\pi k/m)\frac{1-\exp(\mathfrak{i}\pi k)}{1-\exp(\mathfrak{i}\pi k/m)} = 0.$$

This finishes the proof. $\square$

Finally, the following lemma is a refinement of Lemma 1.7 of Tsybakov (2008) regarding the orthogonality of the trigonometric basis on the uniform grid.

Lemma A4.3. Let $m \in \mathbb{N}$ and $k \leq 1, 2, \ldots m-1$.

1. If $l = 1, \ldots, m-1$, then

$$m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) = \delta_{k,l}.$$

2. If $l = m, m+1, \ldots$, then

$$m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) = \begin{cases}\sqrt{2}\mathbb{1}_{l/m\in 2\mathbb{Z}}, & k=1,\\ \mathbb{1}_{(l-k)/m\in 2\mathbb{Z}} + \mathbb{1}_{(l+k)/m\in 2\mathbb{Z}}, & k = 2, 4, \ldots, m-1,\\ \mathbb{1}_{(l-k)/m\in 2\mathbb{Z}} - \mathbb{1}_{(l+k-2)/m\in 2\mathbb{Z}}, & k = 3, 5, \ldots, m-1.\end{cases}$$

Proof of Lemma A4.3. The setting where $l = 1, \ldots, m-1$ is exactly Lemma 1.7 of Tsybakov (2008). For the other setting, we consider a few separate cases. Note that the last line in each of the following cases is a consequence of Lemma A4.2.

1. $k = 1$ and l is even.

$$\begin{aligned} m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) &= \sqrt{2}m^{-1}\sum_{j=1}^{m}\cos(\pi lj/m)\\ &= \frac{1}{\sqrt{2}m}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi lj/m\right) + \exp\left(-\mathfrak{i}\pi lj/m\right)\right)\\ &= \sqrt{2}\mathbb{1}_{l/m\in 2\mathbb{Z}}. \end{aligned}$$

2. $k = 1$ and l is odd.

$$\begin{aligned}
m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) &= \sqrt{2}m^{-1}\sum_{j=1}^{m}\sin(\pi(l-1)j/m) \\
&= \frac{1}{\sqrt{2}m\mathfrak{i}}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l-1)j/m\right) - \exp\left(-\mathfrak{i}\pi(l-1)j/m\right)\right) \\
&= 0.
\end{aligned}$$

3. k, l are both even.

$$\begin{aligned}
&m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) \\
&\quad= m^{-1}\sum_{j=1}^{m}\left(\cos(\pi(l-k)j/m) + \cos(\pi(l+k)j/m)\right) \\
&\quad=\frac{1}{2m}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l-k)j/m\right) + \exp\left(-\mathfrak{i}\pi(l-k)j/m\right)\right) \\
&\qquad+ \frac{1}{2m}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l+k)j/m\right) + \exp\left(-\mathfrak{i}\pi(l+k)j/m\right)\right) \\
&\quad= \mathbb{1}_{(l-k)/m\in 2\mathbb{Z}} + \mathbb{1}_{(l+k)/m\in 2\mathbb{Z}}.
\end{aligned}$$

4. k is even and l is odd.

$$\begin{aligned}
&m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m) \\
&\quad= m^{-1}\sum_{j=1}^{m}\left(\sin(\pi(l-k-1)j/m) + \sin(\pi(l+k-1)j/m)\right) \\
&\quad=\frac{1}{2m\mathfrak{i}}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l-k-1)j/m\right) - \exp\left(-\mathfrak{i}\pi(l-k-1)j/m\right)\right) \\
&\qquad+ \frac{1}{2m\mathfrak{i}}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l+k-1)j/m\right) - \exp\left(-\mathfrak{i}\pi(l+k-1)j/m\right)\right) \\
&\quad= 0.
\end{aligned}$$

5. $k > 1$ is odd and l is even.

$$
\begin{aligned}
&m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m)\\
&\quad= m^{-1}\sum_{j=1}^{m}\left(\sin(\pi(l+k-1)j/m)-\sin(\pi(l-k+1)j/m)\right)\\
&\quad=\frac{1}{2m\mathfrak{i}}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l+k-1)j/m\right)-\exp\left(-\mathfrak{i}\pi(l+k-1)j/m\right)\right)\\
&\qquad-\frac{1}{2m\mathfrak{i}}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l-k+1)j/m\right)-\exp\left(-\mathfrak{i}\pi(l-k+1)j/m\right)\right)\\
&\quad=0.
\end{aligned}
$$

6. $k, l > 1$ are both odd.

$$
\begin{aligned}
&m^{-1}\sum_{j=1}^{m}\varphi_k(j/m)\varphi_l(j/m)\\
&\quad= m^{-1}\sum_{j=1}^{m}\left(\cos(\pi(l-k)j/m)-\cos(\pi(l+k-2)j/m)\right)\\
&\quad=\frac{1}{2m}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l-k)j/m\right)+\exp\left(-\mathfrak{i}\pi(l-k)j/m\right)\right)\\
&\qquad-\frac{1}{2m}\sum_{j=1}^{m}\left(\exp\left(\mathfrak{i}\pi(l+k-2)j/m\right)+\exp\left(-\mathfrak{i}\pi(l+k-2)j/m\right)\right)\\
&\quad=\mathbb{1}_{(l-k)/m\in 2\mathbb{Z}}-\mathbb{1}_{(l+k-2)/m\in 2\mathbb{Z}}.
\end{aligned}
$$

Combining these calculations together proves the claim. □

A.4.2 Proofs for Section 4.2

A.4.2.1 Proofs for Section 4.2.1

The proof of Proposition 4.1 relies on the following two lemmata, which we state for completeness.

Lemma A4.4 (Chao and Strawderman (1972))**.** Let $X \sim \mathrm{Bin}(n,p)$. Then,

$$\mathbb{E}\left(\frac{1}{X+1}\right)=\frac{1-(1-p)^{n+1}}{(n+1)p}.$$

Lemma A4.5 (Boland et al. (2002))**.** Let $Y \sim \mathrm{Bin}(n,p)$ and $X = \sum_{i=1}^n X_i$, where the $X_i \sim \mathrm{Bin}(1,p_i)$ are independent. Then, Y is stochastically smaller than X if and only if $p \leq (\prod_{i=1}^n p_i)^{1/n}$.

Proof of Proposition 4.1. Let F denote the distribution function corresponding to f. Then,

$$
\begin{aligned}
\mathbb{E}_T \tilde{N}_h &= \mathbb{E}_T \sum_{i=1}^n \sum_{j=1}^m \mathbb{1}_{\left\{\exists j' \neq j : t_{i,j'} \in (t_{i,j}, t_{i,j}+h]\right\}} \\
&= \sum_{i=1}^n \sum_{j=1}^m \mathbb{P}\left(\exists j' \neq j : t_{i,j'} \in (t_{i,j}, t_{i,j}+h]\right) \\
&= \sum_{i=1}^n \sum_{j=1}^m \int_0^1 \mathbb{P}\left(\exists j' \neq j : t_{i,j'} \in (t, t+h] | t_{i,j} = t\right) f(t) dt \\
&= \sum_{i=1}^n \sum_{j=1}^m \int_0^1 \left(1 - \mathbb{P}\left(\forall j' \neq j : t_{i,j'} \notin (t, t+h] | t_{i,j} = t\right)\right) f(t) dt \\
&= \sum_{i=1}^n \sum_{j=1}^m \left(1 - \int_0^1 \left(1 - F(t+h) - F(t)\right)^{m-1} f(t) dt\right) \\
&= \sum_{i=1}^n \sum_{j=1}^m \left(1 - \int_0^1 \left(1 - hf(t) + o(h)\right)^{m-1} f(t) dt\right).
\end{aligned}
$$

Under the setting of the first claim, note that

$$\int_0^1 \left(1 - hf(t) + o(h)\right)^{m-1} f(t) dt = 1 - h + o(h).$$

Substituting this into the previous display yields the first claim. For the third claim note that For the third claim, observe that

$$\int_0^1 \left(1 - hf(t) + o(h)\right)^{m-1} f(t) dt \asymp \exp(-mh),$$

which implies that

$$\mathbb{P}\left(\tilde{N}_h \neq N\right) \leq \sum_{i=1}^n \sum_{j=1}^m \int_0^1 \left(1 - hf(t) + o(h)\right)^{m-1} f(t) dt \to 0.$$

This proves the third claim.

For the remaining case, consider a non-homogeneous Poisson process with intensity function $mf(\cdot)$ on $(0,1)$. It is well known that the unordered arrival times have the same distribution as the

$t_{i,j}$. Let M denote the corresponding point process. It is easy to see that

$$\mathbb{P}\left(\exists j' \neq j : t_{i,j'} \in (t, t+h] | t_{i,j} = t\right) \asymp \mathbb{P}\left(M((t, t+h]) \geq 1\right).$$

But, the right hand side of the above display satisfies $\mathbb{P}\left(M((t, t+h]) \geq 1\right) = mf(t)h + o(mh)$ since M is a Poisson process. Since f is bounded from above and below, it follows that

$$\int_0^1 \mathbb{P}\left(\exists j' \neq j : t_{i,j'} \in (t, t+h] | t_{i,j} = t\right) f(t) dt \asymp mh.$$

Finally, it is left to show that $\mathbb{E}(N_h + 1)^{-1} \asymp (nm^2h)^{-1}$. Without the loss of generality, assume that m is even. For $i = 1, \ldots, n$ and $j = 1, \ldots, m/2$, define

$$u_{i,j} \triangleq t_{i,(2j+1)} - t_{i,(2j-1)}$$

with the convention that $t_{i,(2m+1)} = 1$ for all $i = 1, \ldots, n$. Define the following sets of random variables:

$$\begin{aligned}
&W_{i,j} \triangleq \mathbb{1}_{t_{i,(2j)} \in (t_{i,(2j-1)}, t_{i,(2j-1)} + h]}, \\
&X_{i,j} \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \min(ch/u_{i,j}, 1)), \\
&Y_i \sim \text{Bin}(m/2, cmh),
\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m/2$. Then, it follows that

$$N_h = \sum_{i=1}^{n} \sum_{j=1}^{m/2} W_{i,j}.$$

Since $f(\cdot)$ is bounded away from zero and infinity, there exists a constant $c > 0$ such that for all $W_{i,j}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, we have $\mathbb{P}(W_{i,j} | \{u_{\cdot,\cdot}\}) \geq \min(ch/u_{i,j}, 1)$. Moreover, conditioned on $\{u_{\cdot,\cdot}\}$, it is easy to see that $W_{i,j}$ is independent of $W_{i',j'}$ if $(i,j) \neq (i',j')$. Therefore, we have that $X_{i,j} \preceq W_{i,j}$, where $\preceq$ denotes stochastic ordering. Now, by construction, $\sum_{j=1}^{m/2} u_{i,j} \leq 1$, so it follows that

$$chm \leq \left(\prod_{j=1}^{m/2} \min\left(\frac{ch}{u_{i,j}}, 1\right) \right)^{2/m}.$$

Thus, Lemma A4.5 implies that $Y_i \preceq \sum_{j=1}^{m/2} X_{i,j}$. Combining these calculations, we see that

$$\begin{aligned}
\mathbb{E}_T\left(\frac{1}{N_h+1}\right) &= \mathbb{E}_T\left(\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m/2} W_{i,j}+1}\right) \\
&\le \mathbb{E}_T\left(\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m/2} X_{i,j}+1}\right) \\
&\le \mathbb{E}_T\left(\frac{1}{\sum_{i=1}^{n} Y_i+1}\right) \\
&= \frac{1-(1-chm)^{nm/2+1}}{(nm/2+1)mh/2} \\
&\le \frac{1}{nm^2h/2+hm/2},
\end{aligned}$$

where the penultimate line follows from Lemma A4.4. By Jensen's inequality, we have

$$\mathbb{E}_T\left(\frac{1}{N_h+1}\right) \ge \frac{1}{\mathbb{E}_T N_h+1} = \frac{1}{nm^2h/2+1}.$$

Thus,

$$\frac{1}{nm^2h/2+1} \le \mathbb{E}_T\left(\frac{1}{N_h+1}\right) \le \frac{1}{nm^2h/2+mh/2}.$$

Since $mh \to 0$, this finishes the proof. □

A.4.2.2 Proofs for Section 4.2.3

We start by bounding the squared bias of the resultant differenced linear model from equation (4.2.2).

Lemma A4.6. Consider the model given in equation (4.2.1). Assume (4.7) and (4.8). The bias term satisfies

$$\mathbb{E}_T\left(N^{-1}\left\|\boldsymbol{\Delta}^{(\gamma)}\right\|_2^2\right) = \mathcal{O}\left(s_\gamma^* K_\gamma^{-2\alpha} + L^2h^2\right).$$

Proof of Lemma A4.6. Recall that each entry of $\boldsymbol{\Delta}^{(\gamma)}$ may be written as

$$\boldsymbol{\Delta}_{i,j}^{(\gamma)} = \sum_{k=K_\gamma+1}^{\infty} [\varphi_k(t_{i,(j+1)})\mathbf{z}_i(t_{i,(j+1)}) - \varphi_k(t_{i,(j)})\mathbf{z}_i(t_{i,(j)})]^\intercal \daleth_k^* + \xi_i(t_{i,(j+1)}) - \xi_i(t_{i,(j)})$$

for some $(i,j) \in \mathscr{A}_h$. We consider the two parts separately. For the first term, it follows immediately from assumption (4.8) that

$$\mathbb{E}_T\left(\sum_{k=K_\gamma+1}^{\infty}[\varphi_k(t_{i,(j+1)})\mathbf{z}_i(t_{i,(j+1)}) - \varphi_k(t_{i,(j)})\mathbf{z}_i(t_{i,(j)})]^\mathsf{T}\gimel_k^*\right)^2 = \mathcal{O}(s_\gamma^* K_\gamma^{-2\alpha}).$$

By assumption (4.7), we may bound the second term by

$$\left(\xi_i(t_{i,(j+1)}) - \xi_i(t_{i,(j)})\right)^2 \leq L^2h^2.$$

Combining these two bounds finishes the proof. □

Next, we prove the result for the low-dimensional setting.

Proof of Proposition 4.2. Indeed, note that

$$\begin{aligned}\hat{\gimel}^{\text{LD}} &= \left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\left(\boldsymbol{\Psi}\gimel^* + \eta + \boldsymbol{\Delta}^{(\gamma)}\right)\\ &= \gimel^* + \left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\eta} + \left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Delta}^{(\gamma)}.\end{aligned}$$

Bounding each of the two terms separately, we have for the first term that

$$\mathbb{E}\left\|\left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\eta\right\|_2^2 = \operatorname{tr}\left[\left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\right]\sigma_\eta^2 = \mathcal{O}\left(\frac{s_\gamma^* K_\gamma}{N}\right),$$

which follows from assumption (4.1). For the second term, invoking Lemma A4.6 implies that

$$\mathbb{E}_T\left\|\left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Delta}^{(\gamma)}\right\|_2^2 \leq \mathbb{E}_T\left\|\left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T}\right\|_2^2\left\|\boldsymbol{\Delta}^{(\gamma)}\right\|_2^2 = \mathcal{O}\left(s_\gamma^* K_\gamma^{-2\alpha} + L^2h^2\right),$$

which finishes the proof. □

Proof of Theorem 4.3. The first claim follows immediately from Theorem 6.2 of Bühlmann and van de Geer (2011). Then, for the second claim, applying Corollary 6.5 of Bühlmann and van de Geer (2011) yields

$$\left\|\hat{\gimel}^{\text{HD}} - \gimel^*\right\|_2^2 \leq 6\lambda^2 s_\gamma^* K_\gamma\left(\frac{3\mathbb{E}_T(N^{-1}\|\boldsymbol{\Delta}^{(\gamma)}\|_2^2)}{\lambda^2 s_\gamma^* K_\gamma} + \frac{16}{\phi^2_{\text{adap},\Psi}}\right)^2.$$

Now, Lemma A4.6 implies that

$$\left\|\hat{\gimel}^{\mathrm{HD}}-\gimel^{*}\right\|_{2}^{2}=\mathcal{O}\left(\lambda^{2} s_{\gamma}^{*} K_{\gamma}\left(\frac{s_{\gamma}^{*} K_{\gamma}^{-2 \alpha}+L^{2} h^{2}}{\lambda^{2} s_{\gamma}^{*} K_{\gamma}}+\phi_{\mathrm{adap}, \Psi}^{-2}\right)^{2}\right),$$

which finishes the proof. □

A.4.3 Proofs for Section 4.3

A.4.3.1 Proofs for Section 4.3.2

Proof of Proposition 4.4. Throughout this proof, we consider the model given by equation (4.3.2). Recall from equation (4.1.5) that

$$\operatorname{MISE}(\hat{\boldsymbol{\beta}}^{\mathrm{LD}})=\mathbb{E}\mathbb{E}_{T} \sum_{k=1}^{K_{\beta}}\left\|\hat{\beth}_{k}^{\mathrm{LD}}-\beth_{k}^{*}\right\|_{2}^{2}+\mathbb{E}\mathbb{E}_{T} \sum_{k=K_{\beta}+1}^{\infty}\left\|\beth_{k}^{*}\right\|_{2}^{2}.$$

We consider each of the two sums separately. Suppose temporarily that $k \leq K_{\beta}$. Then, the risk in estimating $\beth_{k}^{*}$ by $\hat{\beth}_{k}^{\mathrm{LD}}$ is given by

$$\begin{aligned} \mathbb{E}\mathbb{E}_{T}\left\|\hat{\beth}_{k}^{\mathrm{LD}}-\beth_{k}^{*}\right\|_{2}^{2} &=\mathbb{E}\mathbb{E}_{T}\left\|\left(\mathbf{X}^{\top} \mathbf{X}\right)^{-1} \mathbf{X}^{\top} \boldsymbol{\zeta}_{k}\right\|_{2}^{2} \\ &=\operatorname{tr}\left(\left(\mathbf{X}^{\top} \mathbf{X}\right)^{-1} \mathbf{X}^{\top}\left(\mathbb{E}\mathbb{E}_{T} \boldsymbol{\zeta}_{k} \boldsymbol{\zeta}_{k}^{\top}\right) \mathbf{X}\left(\mathbf{X}^{\top} \mathbf{X}\right)^{-1}\right). \end{aligned}$$

We directly compute $\mathbb{E}\boldsymbol{\zeta}_{k} \boldsymbol{\zeta}_{k}^{\top}$. Note that $\mathbb{E}\zeta_{i,k}=0$ for all $i=1,\ldots,n$. Then, the covariance matrix is diagonal since observations corresponding to different individuals are independent. Therefore, it is left to compute the value of the diagonal entries. Fix $i=1,\ldots,n$ arbitrarily. Then,

$$\begin{aligned} \mathbb{E}\mathbb{E}_{T} \zeta_{i,k}^{2} \leq & 3 \mathbb{E}\mathfrak{o}_{i,k}^{2}+3 m_{i}^{-2} \mathbb{E}\mathbb{E}_{T}\left(\sum_{j=1}^{m_{i}} \varepsilon_{i}(t_{i,j}) \varphi_{k}(t_{i,j})\right)^{2} \\ &+3 m_{i}^{-2} \mathbb{E}\mathbb{E}_{T}\left(\sum_{j=1}^{m_{i}}\left(\mathbf{x}_{i}^{\top} \boldsymbol{\beta}(t_{i,j})+\xi_{i}(t_{i,j})\right) \varphi_{k}(t_{i,j})-\mathbf{x}_{i}^{\top} \beth_{k}^{*}-\mathfrak{o}_{i,k}\right)^{2} \end{aligned}$$

We bound each of the three terms separately. By definition, we have that $\mathbb{E}\mathfrak{o}_{i,k}^{2}=\sigma_{\mathfrak{o},k}^{2}$. Next,

$$m_{i}^{-2} \mathbb{E}\mathbb{E}_{T}\left(\sum_{j=1}^{m_{i}} \varepsilon_{i}(t_{i,j}) \varphi_{k}(t_{i,j})\right)^{2}=m_{i}^{-2} \sum_{j=1}^{m_{i}} \mathbb{E}\varepsilon_{i}^{2}(t_{i,j}) \mathbb{E}_{T} \varphi_{k}^{2}(t_{i,j})=m_{i}^{-1} \sigma_{\varepsilon}^{2}.$$

For the last term, an expansion yields

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_T\left(\sum_{j=1}^{m_i}\left(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}(t_{i,j})+\xi_i(t_{i,j})\right)\varphi_k(t_{i,j})-\mathbf{x}_i^\mathsf{T}\beth_k^*-\mathfrak{o}_{i,k}\right)^2\\
&\quad=\mathbb{E}\mathbb{E}_T\left(\sum_{a=1}^{\infty}\left(\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right)\left(\sum_{j=1}^{m_i}\varphi_a(t_{i,j})\varphi_k(t_{i,j})-\delta_{a,k}\right)\right)^2\\
&\quad=\mathbb{E}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\left(\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right)\left(\mathbf{x}_i^\mathsf{T}\beth_b^*+\mathfrak{o}_{i,b}\right)\\
&\qquad\qquad\times\mathbb{E}_T\left(\sum_{j=1}^{m_i}\varphi_a(t_{i,j})\varphi_k(t_{i,j})-\delta_{a,k}\right)\left(\sum_{j=1}^{m_i}\varphi_b(t_{i,j})\varphi_k(t_{i,j})-\delta_{b,k}\right)
\end{aligned}$$

Applying Lemma A4.1 shows that

$$\begin{aligned}
&\mathbb{E}_T\left(\sum_{j=1}^{m_i}\varphi_a(t_{i,j})\varphi_k(t_{i,j})-\delta_{a,k}\right)\left(\sum_{j=1}^{m_i}\varphi_b(t_{i,j})\varphi_k(t_{i,j})-\delta_{b,k}\right)\\
&\quad\le m_i(\delta_{a+b,2k}+\delta_{a+2k,b}+\delta_{a,b+2k}+2\delta_{a,b}+2\delta_{a+1,b}+2\delta_{a,b+1}+\delta_{a,2k}\delta_{b,1}+\delta_{b,2k}\delta_{a,1}).
\end{aligned}$$

By substitution, we have the following bound

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_T\left(\sum_{j=1}^{m_i}\left(\mathbf{x}_i^\mathsf{T}\beta(t_{i,j})+\xi_i(t_{i,j})\right)\varphi_k(t_{i,j})-\mathbf{x}_i^\mathsf{T}\beth_k^*-\mathfrak{o}_{i,k}\right)^2\\
&\le m_i\mathbb{E}\sum_{a=1}^{2k-1}\left|\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right|\left|\mathbf{x}_i^\mathsf{T}\beth_{2k-a}^*+\mathfrak{o}_{i,2k-a}\right|\\
&\quad+2m_i\mathbb{E}\sum_{a=1}^{\infty}\left|\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right|\left|\mathbf{x}_i^\mathsf{T}\beth_{2k+a}^*+\mathfrak{o}_{i,2k+a}\right|\\
&\quad+2m_i\mathbb{E}\sum_{a=1}^{\infty}\left(\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right)^2\\
&\quad+4m_i\mathbb{E}\sum_{a=1}^{\infty}\left|\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right|\left|\mathbf{x}_i^\mathsf{T}\beth_{a+1}^*+\mathfrak{o}_{i,a+1}\right|\\
&\quad+2m_i\mathbb{E}\left|\mathbf{x}_i^\mathsf{T}\beth_1^*+\mathfrak{o}_{i,1}\right|\left|\mathbf{x}_i^\mathsf{T}\beth_{2k}^*+\mathfrak{o}_{i,2k}\right|.
\end{aligned}$$

From assumptions (4.14) and (4.15) in conjunction with Parseval's Theorem, it follows that

$$2m_i\mathbb{E}\sum_{a=1}^{\infty}\left(\mathbf{x}_i^\mathsf{T}\beth_a^*+\mathfrak{o}_{i,a}\right)^2=\mathcal{O}(m_ig(n)).$$

Now, using the inequality $2uv \leq u^2 + v^2$ and the above, we have that

$$
\begin{aligned}
&2m_i \mathbb{E} \sum_{a=1}^{\infty} \left| \mathbf{x}_i^{\mathsf{T}} \beth_a^* + \mathfrak{o}_{i,a} \right| \left| \mathbf{x}_i^{\mathsf{T}} \beth_{2k+a}^* + \mathfrak{o}_{i,2k+a} \right| \\
&\qquad \leq m_i \mathbb{E} \sum_{a=1}^{\infty} \left(\mathbf{x}_i^{\mathsf{T}} \beth_a^* + \mathfrak{o}_{i,a} \right)^2 + m_i \mathbb{E} \sum_{a=1}^{\infty} \left(\mathbf{x}_i^{\mathsf{T}} \beth_{2k+a}^* + \mathfrak{o}_{i,2k+a} \right)^2 \\
&\qquad = \mathcal{O}(m_i g(n)).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
m_i \mathbb{E} \sum_{a=1}^{2k-1} \left| \mathbf{x}_i^{\mathsf{T}} \beth_a^* + \mathfrak{o}_{i,a} \right| \left| \mathbf{x}_i^{\mathsf{T}} \beth_{2k-a}^* + \mathfrak{o}_{i,2k-a} \right| &= \mathcal{O}(m_i g(n)), \\
4 m_i \mathbb{E} \sum_{a=1}^{\infty} \left| \mathbf{x}_i^{\mathsf{T}} \beth_a^* + \mathfrak{o}_{i,a} \right| \left| \mathbf{x}_i^{\mathsf{T}} \beth_{a+1}^* + \mathfrak{o}_{i,a+1} \right| &= \mathcal{O}(m_i g(n)).
\end{aligned}
$$

Thus, combining all the results yields

$$
\mathbb{E}\mathbb{E}_T \left\| \hat{\beth}_k^{\mathrm{LD}} - \beth_k^* \right\|_2^2 = \operatorname{tr}\left(\left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathsf{T}} \left(\sigma_{\mathfrak{o},k}^2 I_n + \mathcal{O}(g(n)) \mathbf{M}^{-1} \right) \mathbf{X} \left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} \right),
$$

where $M \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose i'th entry is m_i. Therefore,

$$
\sum_{k=1}^{K_\beta} \left\| \hat{\beth}_k^{\mathrm{LD}} - \beth_k^* \right\|_2^2 = \operatorname{tr}\left(\left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathsf{T}} \left(\mathcal{O}\left(1\right) \mathbf{I}_n + \mathcal{O}(g(n) K_\beta) \mathbf{M}^{-1} \right) \mathbf{X} \left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} \right),
$$

since $\sum_{k=1}^{K_\beta} \sigma_{\mathfrak{o},k}^2 = \mathcal{O}(1)$. Recalling that

$$
\sum_{k=K_\beta+1}^{\infty} \|\beth_k^*\|_2^2 = \mathcal{O}(s_\beta^* K_\beta^{-2\alpha})
$$

finishes the proof. $\square$

Lemma A4.7. Consider the model from equation (4.3.2). Assume (4.6), (4.14), and (4.16). Then, $\zeta_{i,k} \sim \mathcal{SG}(\varsigma_{\zeta,i,k}^2)$ with respect to the joint probability measure on $\varepsilon_i(t_{i,j})$, $\xi_i(\cdot)$, and $t_{i,j}$, where $\varsigma_{\zeta,i,k}^2 = \mathcal{O}(\varsigma_{\mathfrak{o},k}^2 + g(n) m_i^{-1})$.

Proof of Lemma A4.7. Consider first the model from equation (4.3.2). We partition $\zeta_{i,k}$ into four terms and show each term is sub-Gaussian.

$$\zeta_{i,k} = \underbrace{\mathfrak{o}_{i,k}}_{(I)} + \underbrace{m_i^{-1}\sum_{j=1}^{m_i}\varepsilon_i(t_{i,j})\varphi_k(t_{i,j})}_{(II)} + \underbrace{m_i^{-1}\sum_{j=1}^{m_i}\mathbf{x}_i^{\mathsf{T}}\left(\beta(t_{i,j})\varphi_k(t_{i,j}) - \daleth_k^*\right)}_{(III)}$$
$$+ \underbrace{m_i^{-1}\sum_{j=1}^{m_i}\left(\xi_i(t_{i,j})\varphi_k(t_{i,j}) - \mathfrak{o}_{i,k}\right)}_{(IV)}.$$

For the first term, by assumption (4.16), $\mathfrak{o}_{i,k} \sim \mathcal{SG}(\varsigma_{\mathfrak{o},k}^2)$. Next, we have, for any fixed $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\mathbb{E}_T \exp\left(\lambda m_i^{-1}\sum_{j=1}^{m_i}\varepsilon_{i,j}\varphi_k(t_{i,j})\right) &= \prod_{j=1}^{m_i}\mathbb{E}\mathbb{E}_T \exp\left(\lambda m_i^{-1}\varepsilon_{i,j}\varphi_k(t_{i,j})\right) \\
&= \prod_{j=1}^{m_i}\mathbb{E}_T \exp\left(\frac{\varsigma_\varepsilon^2\lambda^2\varphi_k^2(t_{i,j})}{2m_i^2}\right) \\
&\le \exp\left(\frac{2\varsigma_\varepsilon^2 m_i^{-1}\lambda^2}{2}\right).
\end{aligned}$$

In the second equality, we have used assumption (4.6). Thus, $(II) \sim \mathcal{SG}(2\varsigma_\varepsilon^2 m_i^{-1})$. Then, for the third term, we see that for $\lambda > 0$,

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_T \exp\left(\lambda m_i^{-1}\sum_{j=1}^{m_i}\mathbf{x}_i^{\mathsf{T}}\left(\beta(t_{i,j})\varphi_k(t_{i,j}) - \daleth_k^*\right)\right) \\
&\le \mathbb{E}\prod_{j=1}^{m_i}\mathbb{E}_T \exp\left(\lambda m_i^{-1}\mathbf{x}_i^{\mathsf{T}}\left(\beta(t_{i,j})\varphi_k(t_{i,j}) - \daleth_k^*\right)\right) \\
&\le \mathbb{E}\prod_{j=1}^{m_i}\exp\left(\frac{\mathcal{O}(g(n))m_i^{-2}\lambda^2}{2}\right) \\
&\le \exp\left(\frac{\mathcal{O}(g(n)m_i^{-1})\lambda_2^2}{2}\right).
\end{aligned}$$

Hence, $m_i^{-1}\sum_{j=1}^{m_i}\mathbf{x}_i^{\mathsf{T}}\left(\beta(t_{i,j})\varphi_k(t_{i,j}) - \daleth_k^*\right) \sim \mathcal{SG}(\mathcal{O}(g(n)m_i^{-1})$. Finally, from Lemma 1.8 of Tsybakov (2008), assumption (4.16) implies that $(\xi_i(\cdot))_{i=1}^n$ are uniformly bounded by a constant, which we temporarily denote by $c > 0$. Then, by an analogous argument as above, it follows that

$$m_i^{-1}\sum_{j=1}^{m_i}\left(\xi_i(t_{i,j})\varphi_k(t_{i,j}) - \mathfrak{o}_{i,k}\right) \sim \mathcal{SG}(c^2 m_i^{-1}).$$

Combining these results finishes the proof. $\square$

Proof of Theorem 4.5. We proceed by modifying the standard lasso arguments to account for the different sub-Gaussian parameters of the noise term. From the Basic Inequality (Lemma 6.1 of Bühlmann and van de Geer (2011)), it follows that

$$n^{-1}\left\|\mathbf{X}\left(\hat{\daleth}_k-\daleth_k^*\right)\right\|_2^2+\lambda_k\left\|\hat{\daleth}_k\right\|_1\leq 2n^{-1}\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}\left(\hat{\daleth}_k-\daleth_k^*\right)+\lambda_k\left\|\daleth_k^*\right\|_1.$$

To bound the first term on the right hand side, we similarly apply an $\ell_1-\ell_\infty$ bound to obtain

$$2n^{-1}\left|\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}\left(\hat{\daleth}_k-\daleth_k^*\right)\right|\leq 2n^{-1}\max_{j=1,\ldots,p}\left|\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}_j\right|\left\|\hat{\daleth}_k-\daleth_k^*\right\|_1.$$

Define the set $\mathscr{T}_k$ as

$$\mathscr{T}_k\triangleq\left\{2n^{-1}\max_{j=1,\ldots,p}\left|\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}_j\right|\leq\lambda_{0,k}\right\}.$$

Then, for any value of $r>0$,

$$\begin{aligned}\mathbb{P}\left(\mathscr{T}_k^{\mathsf{c}}\right)&\leq 2\sum_{j=1}^{p}\mathbb{P}\left(2n^{-1}\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}_j>\lambda_{0,k}\right)\\&\leq 2\sum_{j=1}^{p}\mathbb{P}\left(\exp\left(r\boldsymbol{\zeta}_k^{\mathsf{T}}X_j\right)>\exp\left(\frac{rn\lambda_{0,k}}{2}\right)\right)\\&\leq 2\sum_{j=1}^{p}\exp\left(-\frac{rn\lambda_{0,k}}{2}\right)\mathbb{E}\exp\left(r\boldsymbol{\zeta}_k^{\mathsf{T}}\mathbf{X}_j\right)\\&\leq 2\sum_{j=1}^{p}\exp\left(-\frac{rn\lambda_{0,k}}{2}\right)\exp\left(\frac{r^2}{2}\sum_{i=1}^{n}\varsigma_{\zeta,i,k}^2\mathbf{x}_{i,j}^2\right)\end{aligned}$$

Since the value of $r>0$ was arbitrary, setting

$$r=\left(\sum_{i=1}^{n}\varsigma_{\eta,i,k}^2\mathbf{x}_{i,j}^2\right)^{-1}\frac{n\lambda_{0,k}}{2}$$

yields the bound

$$\mathbb{P}\left(\mathscr{T}^{\mathsf{c}}\right)\leq 2p\exp\left(-\left(t^2/2+\log(p)\right)\right)\leq 2\exp\left(-t^2/2\right).$$

Thus, we may restrict our attention to the event $\mathscr{T}_k$. The desired results follow by applying Theorem 6.1 and Corollary 6.5 of Bühlmann and van de Geer (2011) respectively. $\square$

A.4.3.2 Proofs for Section 4.3.3

Lemma A4.8. Assume (4.10) and (4.14). Then,

$$\sum_{k=1}^{m-1} \|\daleth_k\|_2^2 = \mathcal{O}(s_\beta^* m^{-2\alpha}).$$

Proof of Lemma A4.8. Indeed, by the second half of Lemma A4.3, it follows that

$$\daleth_k = \begin{cases} \sqrt{2}\sum_{r=1}^\infty \gimel_{2rm}^*, & \text{if } k=1, \\ \sum_{r=1}^\infty \gimel_{2rm+k}^* + \gimel_{2rm-k}^*, & \text{if } k=2,4,\ldots,m-1, \\ \sum_{r=1}^\infty \gimel_{2rm+k}^* - \gimel_{2rm+2-k}^*, & \text{if } k=3,5,\ldots,m-1. \end{cases}$$

Define the following sequence of constants $(a_k)_{k=1}^\infty$ from Tsybakov (2008)

$$a_k = \begin{cases} k^\alpha, & \text{for even } k, \\ (k-1)^\alpha, & \text{for odd } k. \end{cases}$$

Since $\alpha > 1/2$, let $c > 0$ be a constant such that $\sum_{r=1}^\infty r^{-2\alpha} \leq c$. Now, for $k=1$, it follows that

$$\|\daleth_k\|_2^2 \leq 2\left(\sum_{r=1}^\infty a_{2rm}^2 \|\gimel_{2rm}^*\|_2^2\right)\left(\sum_{r=1}^\infty a_{2rm}^{-2}\right) \leq 2cm^{-2\alpha}\left(\sum_{r=1}^\infty a_{2rm}^2\|\gimel_{2rm}^*\|_2^2\right).$$

Similarly, for $k=2,\ldots,m-1$ and $k=3,\ldots,m-1$, we have

$$\|\daleth_k\|_2^2 \leq 2cm^{-2\alpha}\sum_{r=1}^\infty \left(a_{2rm+k}^2\|\gimel_{2rm+k}^*\|_2^2 + a_{2rm-k}^2\|\gimel_{2rm-k}^*\|_2^2\right)$$

and

$$\|\daleth_k\|_2^2 \leq 2cm^{-2\alpha}\sum_{r=1}^\infty \left(a_{2rm+k}^2\|\gimel_{2rm+k}^*\|_2^2 + a_{2rm+2-k}^2\|\gimel_{2rm+2-k}^*\|_2^2\right)$$

respectively. Thus, combining the above calculations yields

$$\sum_{k=1}^{m-1}\|\daleth_k\|_2^2 \leq 2cm^{-2\alpha}\sum_{r=m}^\infty a_r^2\|\gimel_r^*\|_2^2 = \mathcal{O}(s_\beta^* m^{-2\alpha}),$$

which finishes the proof. $\square$

Proof of Proposition 4.6. Note that MISE is given by

$$
\begin{aligned}
\operatorname{MISE}(\hat{\boldsymbol{\beta}}^{\mathrm{LD}}) &= \sum_{k=1}^{K_\beta} \mathbb{E}\|\hat{\gimel}_k^{\mathrm{LD}} - \gimel_k^*\|_2^2 + \sum_{k=K_\beta+1}^{\infty} \|\gimel_k^*\|_2^2 \\
&\leq 2\sum_{k=1}^{K_\beta} \left(\|\daleth_k\|_2^2 + \mathbb{E}\|(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\zeta}_k\|_2^2\right) + \sum_{k=K_\beta+1}^{\infty} \|\gimel_k^*\|_2^2.
\end{aligned}
$$

For the variance term, we have that

$$
\mathbb{E}\|(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\zeta}_k\|_2^2 = \operatorname{tr}\left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbb{E}(\boldsymbol{\zeta}_k\boldsymbol{\zeta}_k^\mathsf{T})\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right).
$$

By independence, for $1 \leq i < j \leq n$, it follows that

$$
\mathbb{E}(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\mathsf{T})_{i,j} = 0.
$$

Thus, $\mathbb{E}(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\mathsf{T})$ is a diagonal matrix. For $i = 1, \ldots, n$ and $k = 1$,

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\mathsf{T})_{i,i} &= \mathbb{E}\left(\mathfrak{o}_{i,k} + m^{-1}\sum_{j=1}^{m}\varphi_k(t_{i,j})\varepsilon_i(t_{i,j}) + \sqrt{2}\sum_{r=1}^{\infty}\mathfrak{o}_{i,2rm}\right)^2 \\
&= \mathbb{E}\left(\mathfrak{o}_{i,k} + \sqrt{2}\sum_{r=1}^{\infty}\mathfrak{o}_{i,2rm}\right)^2 + \sigma_\varepsilon^2 m^{-1} \\
&\leq 2c\left(a_k^2\mathbb{E}\mathfrak{o}_{i,k}^2 + \sum_{r=1}^{\infty}a_{2rm}^2\mathbb{E}\mathfrak{o}_{i,2rm}^2\right) + \sigma_\varepsilon^2 m^{-1}.
\end{aligned}
$$

By performing similar calculations when $k = 2, \ldots, m-1$ and $k = 3, \ldots, m-1$, we have that

$$
\mathbb{E}(\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\mathsf{T})_{i,i} \leq \begin{cases}
2c\left(a_k^2\mathbb{E}\mathfrak{o}_{i,k}^2 + \sum_{r=1}^{\infty}a_{2rm}^2\mathbb{E}\mathfrak{o}_{i,2rm}^2\right) + \sigma_\varepsilon^2 m^{-1}, & k = 1, \\
2c\sum_{r=1}^{\infty}\left(a_{2rm+k}^2\mathbb{E}\mathfrak{o}_{i,2rm+k}^2 + a_{2rm-k}^2\mathbb{E}\mathfrak{o}_{i,2rm}^2\right) + \sigma_\varepsilon^2 m^{-1}, & k \text{ even} \\
2c\sum_{r=1}^{\infty}\left(a_{2rm+k}^2\mathbb{E}\mathfrak{o}_{i,2rm+k}^2 + a_{2rm+2-k}^2\mathbb{E}\mathfrak{o}_{i,2rm+2-k}^2\right) + \sigma_\varepsilon^2 m^{-1}, & \text{otherwise.}
\end{cases}
$$

Since $\xi_i(\cdot) \in \mathcal{W}^{\mathrm{per}}(\alpha, R)$ almost surely by assumption (4.15), it follows from Proposition 1.14 of Tsybakov (2008) that

$$
\sum_{r=1}^{\infty} a_k^2\mathbb{E}\mathfrak{o}_{i,r}^2 = \mathcal{O}(1).
$$

Thus,

$$\sum_{k=1}^{K_\beta} \mathbb{E}(\boldsymbol{\eta}_k \boldsymbol{\eta}_k^\top)_{i,i} = \mathcal{O}(1 + K_\beta m^{-1}).$$

Hence,

$$\sum_{k=1}^{K_\beta} \mathbb{E}\|(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\eta}_k\|_2^2 = \mathcal{O}(1 + K_\beta m^{-1}) \mathrm{tr}[(X^\top X)^{-1}].$$

Invoking Lemma A4.8 and recalling that

$$\sum_{k=K_\beta+1}^{\infty} \|\beth_k^*\|_2^2 = \mathcal{O}(s_\beta^* K_\beta^{-2\alpha})$$

finishes the proof. $\square$

Proof of Theorem 4.7. We start by showing that $\zeta_{i,k}$ is sub-Gaussian with parameter $c_k + m^{-1}\varsigma_\varepsilon^2$. By the second half of Lemma A4.3, we have that

$$\zeta_{i,k} = \begin{cases} \mathfrak{o}_{i,k} + m^{-1}\sum_{j=1}^m \varphi_k(t_{i,j})\varepsilon_i(t_{i,j}) + \sqrt{2}\sum_{r=1}^\infty \mathfrak{o}_{i,2rm}, & k = 1, \\ \mathfrak{o}_{i,k} + m^{-1}\sum_{j=1}^m \varphi_k(t_{i,j})\varepsilon_i(t_{i,j}) + \sum_{r=1}^\infty \mathfrak{o}_{i,2rm+k} + \mathfrak{o}_{i,2rm-k}, & k = 2, 4, \dots, m-1, \\ \mathfrak{o}_{i,k} + m^{-1}\sum_{j=1}^m \varphi_k(t_{i,j})\varepsilon_i(t_{i,j}) + \sum_{r=1}^\infty \mathfrak{o}_{i,2rm+k} - \mathfrak{o}_{i,2rm+2-k}, & k = 3, 5, \dots, m-1. \end{cases}$$

By the first half of Lemma A4.3, it is easy to see that

$$m^{-1}\sum_{j=1}^m \varphi_k(t_{i,j})\varepsilon_i(t_{i,j}) \sim \mathcal{SG}(m^{-1}\varsigma_\varepsilon^2).$$

Then, by the triangle inequality and assumption (4.16), it follows that

$$\zeta_{i,k} \sim \mathcal{SG}(c_k + m^{-1}\varsigma_\varepsilon^2).$$

Next, the ISE can be bounded by

$$\begin{aligned}
\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) &= \sum_{k=1}^{K_\beta} \|\hat{\beth}_k^{\mathrm{HD}} - \beth_k^*\|_2^2 + \sum_{k=K_\beta+1}^{\infty} \|\beth_k^*\|_2^2 \\
&\leq 2\sum_{k=1}^{K_\beta} \|\hat{\beth}_k^{\mathrm{HD}} - \beth_k^* - \daleth_k\|_2^2 + 2\sum_{k=1}^{K_\beta} \|\daleth_k\|_2^2 + \sum_{k=K_\beta+1}^{\infty} \|\beth_k^*\|_2^2 \\
&\leq 2\sum_{k=1}^{K_\beta} \|\hat{\beth}_k^{\mathrm{HD}} - \beth_k^* - \daleth_k\|_2^2 + \mathcal{O}(s_\beta^* m^{-2\alpha} + s_\beta^* K_\beta^{-2\alpha}),
\end{aligned}$$

where we have used Lemma A4.8 in the last line. Now, by Corollary 6.5 of Bühlmann and van de Geer (2011), it follows that, with probability at least $1 - 2\exp(-t^2/2)$,

$$\|\hat{\beth}_k^{\mathrm{HD}} - \beth_k^* - \daleth_k\|_2^2 = \mathcal{O}(s_\beta^* \lambda_k^2).$$

By assumption (4.16),

$$\sum_{k=1}^{K_\beta} c_k = \mathcal{O}(1).$$

Assuming $\lambda_k = 2\lambda_{0,k}$, we have that

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) = \mathcal{O}\left(\frac{s_\beta^* \log(p)}{n} + \frac{s_\beta^* K_\beta \log(p)}{mn} + s_\beta^* m^{-2\alpha} + s_\beta^* K_\beta^{-2\alpha}\right).$$

Choosing $K_\beta \asymp (mn/\log(p))^{1/(2\alpha+1)}$,

$$\mathrm{ISE}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) = \mathcal{O}\left(\frac{s_\beta^* \log(p)}{n} + s_\beta^* \left(\frac{\log(p)}{mn}\right)^{2\alpha/(2\alpha+1)} + s_\beta^* m^{-2\alpha}\right).$$

$\square$

A.4.4 Proofs for Section 4.5

Proof of Theorem 4.8. Indeed, for each $k = 1, \dots, K_\beta$, we may rewrite $\hat{\beth}_k^{\text{DB}}$ as

$$\begin{aligned}\sqrt{n\sigma_{\zeta,k}^{-2}}\left(\hat{\beth}_k^{\text{DB}} - \beth_k^*\right) =& \sqrt{n\sigma_{\zeta,k}^{-2}}\left(\hat{\beth}_k - \beth_k^* + \hat{\boldsymbol{\Theta}}\mathbf{X}^\mathsf{T}\left(\mathbf{X}\beth_k^* - \mathbf{X}\hat{\beth}_k^{\text{HD}} + \boldsymbol{\zeta}_k\right)/n\right) \\ =& \sqrt{n\sigma_{\zeta,k}^{-2}}\left(\mathbf{I}_p - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\right)\left(\hat{\beth}_k^{\text{HD}} - \beth_k^*\right) \\ &+ \sqrt{n\sigma_{\zeta,k}^{-2}}\hat{\boldsymbol{\Theta}}\mathbf{X}^\mathsf{T}\boldsymbol{\zeta}_k/n.\end{aligned}$$

Now, since $\hat{\beth}_{k,1}^{\text{DB}}$ is the first entry of $\hat{\beth}_k^{\text{DB}}$, we set W_k to be the first entry of $\sqrt{n\sigma_{\zeta,k}^{-2}}\hat{\boldsymbol{\Theta}}\mathbf{X}^\mathsf{T}\boldsymbol{\zeta}_k/n$ and Δ_k to be the first entry of $\sqrt{n\sigma_{\zeta,k}^{-2}}\left(\mathbf{I}_p - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\right)\left(\hat{\beth}_k^{\text{HD}} - \beth_k^*\right)$. The first claim follows by Proposition 2.1 of Chernozhukov et al. (2017). Then, an $\ell_1 - \ell_\infty$ bound yields

$$\sup_{k=1,\dots,K_\beta} |\Delta_k| \leq \sqrt{n}\left\|\mathbf{I}_p - \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\right\|_\infty \sup_{k=1,\dots,K_\beta}\left\|\hat{\beth}_k^{\text{HD}} - \beth_k^*\right\|_1.$$

For the other term, Theorem 4.5 implies, with probability at least $1 - 2\exp(-\log^2(p)/2 + \log(K_\beta)) \to 1$, that

$$\sup_{k=1,\dots,K_\beta}\left\|\hat{\beth}_k^{\text{HD}} - \beth_k^*\right\|_1 \leq 4s_\beta^*/\phi_{\text{cc},X}^2 \sup_{k=1,\dots,K_\beta} \lambda_k = \mathcal{O}\left(s_\beta^*\sqrt{\log(p)/n}\right).$$

Combining these bounds, we see that

$$\sup_{k=1,\dots,K_\beta} |\Delta_k| = \mathcal{O}_\mathbb{P}(s_\beta^* \log(p)/\sqrt{n}) = o_\mathbb{P}(1),$$

which finishes the proof. □

Proof of Proposition 4.9. Recall the decomposition for $\beta_1(\cdot)$ as

$$\boldsymbol{\beta}_1(\cdot) = \underline{\boldsymbol{\beta}}_1(\cdot) + \overline{\boldsymbol{\beta}}_1(\cdot).$$

Note that the event

$$\begin{aligned}&\{\forall t \in (0,1) : l(t) \leq \underline{\boldsymbol{\beta}}_1(t) \leq u(t)\} \cap \left\{\forall t \in (0,1) : |\overline{\boldsymbol{\beta}}_1(t)| \leq \delta\right\} \\ &\subseteq \left\{\forall t \in (0,1) : l_\delta(t) \leq \underline{\boldsymbol{\beta}}_1(t) \leq u_\delta(t)\right\}.\end{aligned}$$

For the high-frequency signal, we have under assumption (4.14) that

$$\left|\overline{\boldsymbol{\beta}}(t)\right| = \mathcal{O}\left(K_\beta^{-\alpha}\log(K_\beta)\right)$$

from Chapter 1.21 of Jackson (1941) and Section 87 of Achieser (1992). Since $\delta \asymp K_\beta^{-\alpha}\log(K_\beta)$, for n sufficiently large, the event $\left\{\forall t \in (0,1) : |\overline{\boldsymbol{\beta}}_1(t)| \leq \delta\right\}$ occurs with probability one. Moreover, for n sufficiently large, Theorem 4.8 implies that

$$\mathbb{P}\left(\forall t \in (0,1) : l(t) \leq \underline{\boldsymbol{\beta}}_1(t) \leq u(t)\right) \geq 1-\tau.$$

This proves the first claim. For the second claim, note that

$$\begin{aligned}\sup_{t\in(0,1)} |u_\delta(t) - l_\delta(t)| &= \sup_{t\in(0,1)} \sum_{k=1}^{K_\beta} (b_k - a_k + 2\delta)\,|\varphi_k(t)| \\ &\leq \sqrt{2}\left(\sum_{k=1}^{K_\beta}(b_k - a_k) + 2K_\beta\delta\right).\end{aligned}$$

Now, since $z_{\tau/K_\beta} = \mathcal{O}(\sqrt{\log(K_\beta)})$, assumption (4.18) implies that

$$\sum_{k=1}^{K_\beta}(b_k - a_k) = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\log(K_\beta)/n} + K_\beta\sqrt{g(n)\log(K_\beta)/(nm)}\right).$$

Moreover, note that $\log(K_\beta) \leq \log(n)$. Thus, we have that

$$\sup_{t\in(0,1)} |u_\delta(t) - l_\delta(t)| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\log(n)/n} + K_\beta\sqrt{g(n)\log(n)/(nm)} + K_\beta^{-\alpha}\log(n)\right).$$

Substituting the choice of K_β finishes the proof. □

A.4.5 Yeast Cell Cycle Data

In this section, we apply our methodology to analyze transcription factors affecting the cell cycle of yeast. Versions of this data was previously analyzed in the high-dimensional varying coefficients framework by Wei et al. (2011) and Bai et al. (2019), to which we refer the interested reader for a more detailed description of the data. For our analysis, we use the data from Bai et al. (2019), which consists of $n = 47$ genes and $p = 96$ transcription factors. The response y is the mRNA level, measured seven minutes apart for 119 minutes, yielding $m = 18$ time points for each gene. Since the time points are evenly spaced, we set $t_{i,j} = j/m$. There are no time varying covariates in this data. The model that we consider is

$$y_i(t_{i,j}) = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}(t_{i,j}) + \xi_i(t_{i,j}) + \varepsilon_i(t_{i,j}).$$

This is analogous to the model fit by Bai et al. (2019), who instead combine the noise $\xi_i(t_{i,j}) + \varepsilon_i(t_{i,j})$ and assume an AR(1) covariance structure. In Figure A.4.1, we provide marginal confidence bands for two selected transcription factors: ABF1 and MAC1.

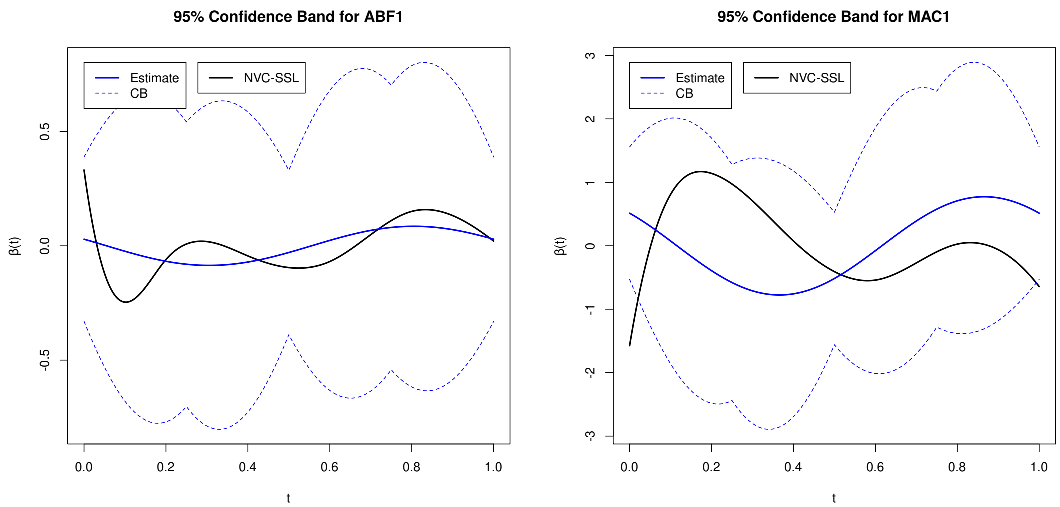

Figure A.4.1: Marginal confidence bands for ABF1 and MAC1. “Estimate” and “CB” are the estimate and confidence band from Section 4.5 respectively while “NVC-SSL” is the methodology from Bai et al. (2019).

From the plot for ABF1, the estimated function seems to follow the general shape of the estimate from Bai et al. (2019), with both estimates entirely contained within the confidence bands. For MAC1, we observe that the estimated curves differ for the two estimation procedures, but the confidence band contains the majority of the estimated curve by NVC-SSL, with the discrepancies at the two endpoints. We remark that the performance of our confidence bands may be anomalous at the boundary points since the theory in Section 4.3 assumes that the varying coefficients are periodic Sobolev functions. In the setting where coefficients are not periodic, convergence still holds on the interior of the interval.

A.4.6 Additional Simulation Results

In this section, we provide the results of the simulations from Section 4.6.

Table A.4.1: Simulations for $\beta(\cdot)$ with Trigonometric Basis

Simulations for $\beta(\cdot)$ with Trigonometric

		s_β^*	15	25	15	25	15	25	15	25
		n	200	200	500	500	200	200	500	500
		$t_{i,j}$	ind	ind	ind	ind	com	com	com	com
Basis	Average Loss	$m = 25$	0.395	0.582	0.195	0.246	0.150	0.252	0.065	0.086
		$m = 50$	0.242	0.368	0.118	0.149	0.091	0.153	0.038	0.051
		$m = 75$	0.182	0.276	0.089	0.114	0.068	0.114	0.028	0.038
		$m = 150$	0.116	0.180	0.054	0.068	0.042	0.071	0.017	0.023
	Average Coverage	$m = 25$	0.685	0.625	0.810	0.815	0.965	0.945	0.970	0.975
		$m = 50$	0.830	0.780	0.940	0.940	0.975	0.985	0.990	0.985
		$m = 75$	0.885	0.850	0.950	0.945	0.970	0.960	0.970	0.985
		$m = 150$	0.925	0.920	0.945	0.965	0.990	0.995	0.985	0.980
	Average Length	$m = 25$	1.368	1.333	1.012	0.933	1.098	1.024	0.928	0.846
		$m = 50$	1.266	1.124	0.963	0.844	1.042	0.969	0.831	0.795
		$m = 75$	1.159	1.001	0.968	0.873	0.986	0.887	0.770	0.694
		$m = 150$	1.019	0.869	0.841	0.734	0.868	0.789	0.681	0.644

Table A.4.2: Simulations for $\beta(\cdot)$ with B-Spline Basis

Simulations for $\beta(\cdot)$ with B-Spline

		s_β^*	15	25	15	25	15	25	15	25
		n	200	200	500	500	200	200	500	500
		$t_{i,j}$	ind	ind	ind	ind	com	com	com	com
Basis	Average Loss	$m = 25$	1.492	2.140	0.738	0.940	1.031	1.645	0.443	0.570
		$m = 50$	1.257	1.883	0.584	0.764	0.970	1.551	0.383	0.506
		$m = 75$	1.158	1.755	0.522	0.682	0.930	1.490	0.372	0.489
		$m = 150$	1.031	1.625	0.432	0.580	0.889	1.451	0.337	0.452
	Average Coverage	$m = 25$	0.870	0.755	0.945	0.945	1.000	1.000	1.000	1.000
		$m = 50$	0.930	0.910	0.980	0.985	1.000	1.000	1.000	1.000
		$m = 75$	0.990	0.970	0.990	1.000	1.000	1.000	1.000	1.000
		$m = 150$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Average Length	$m = 25$	3.755	3.296	3.449	3.050	3.751	3.809	2.870	2.868
		$m = 50$	3.923	3.645	3.709	3.406	4.013	4.089	2.915	2.904
		$m = 75$	4.587	4.208	3.583	3.447	4.407	4.554	3.130	3.038
		$m = 150$	5.155	4.863	4.096	4.015	4.709	4.871	3.309	3.296

Table A.4.3: Simulations for $\gamma(\cdot)$

Simulations for $\gamma(\cdot)$

	s_γ^*	15	25	15	25	15	25	15	25
	$t_{i,j}$	ind	ind	com	com	ind	ind	com	com
	diff type	A	A	A	A	B	B	B	B
Trigonometric	$m = 25$	1.4526	2.1085	0.3633	0.4806	0.7348	0.9096	1.3686	1.5289
Splines	$m = 25$	3.0488	4.0465	0.9057	1.1462	1.2152	1.4874	1.5288	1.7016
Trigonometric	$m = 50$	0.4231	0.5647	0.2007	0.2598	0.3992	0.4844	1.9385	1.9891
Splines	$m = 50$	1.0675	1.3570	0.5743	0.7048	0.6556	0.7819	2.0637	2.1226
Trigonometric	$m = 75$	0.2453	0.3167	0.1493	0.1899	0.3692	0.4238	2.0780	2.1017
Splines	$m = 75$	0.6807	0.8494	0.4437	0.5311	0.5615	0.6430	2.1772	2.2064

APPENDIX 4

Appendix of Chapter 5

A.4.6 Proofs

Proof of Lemma 5.1. By definition of $\hat{\mathbf{\Theta}}_{L_0}$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\text{HS}})^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \mathbf{X}_i, \mathbf{\Theta}^*\rangle_{\text{HS}})^2,$$

which implies that

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2 \leq \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}.$$

If $n^{-1}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2 = 0$, then the result follows. Therefore, we only consider the case where $n^{-1}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2 > 0$. Dividing both sides of the above display by $(n^{-1}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2)^{1/2}$ yields

$$\begin{aligned}\Big(\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2\Big)^{1/2} &\leq \Big(\frac{4}{n}\Big)^{1/2}\frac{\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}}{(\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2)^{1/2}} \\ &\leq \Big(\frac{4}{n}\Big)^{1/2}\sup_{\substack{\mathbf{M}\in\mathbb{R}^{d_1\times d_2}\\ \text{rank}(\mathbf{M})\leq 2r\\ \sum_{i=1}^{n}\langle \mathbf{X}_i,\mathbf{M}\rangle_{\text{HS}}^2>0}}\frac{\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \mathbf{M}\rangle_{\text{HS}}}{(\sum_{i=1}^{n}\langle \mathbf{X}_i, \mathbf{M}\rangle_{\text{HS}}^2)^{1/2}}.\end{aligned}$$

The second inequality follows from the fact that $\text{rank}(\hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*) \leq 2r$. Now, for any $\mathbf{M}$ satisfying the above, there exist matrices $\mathbf{U} \in \mathbb{R}^{d_1\times 2r}$ and $\mathbf{V} \in \mathbb{R}^{d_2\times 2r}$ such that $\mathbf{M} = \mathbf{U}\mathbf{V}^{\mathsf{T}}$. Note that $\langle \mathbf{X}_i, \mathbf{U}\mathbf{V}^{\mathsf{T}}\rangle_{\text{HS}} = \langle \mathbf{X}_i\mathbf{V}, \mathbf{U}\rangle_{\text{HS}}$. Let $\mathbf{X}_{\mathbf{V}} \in \mathbb{R}^{n\times rd_1}$ be the matrix whose ith row is $\text{vec}(\mathbf{X}_i\mathbf{V})$ and $\gamma_{\mathbf{U}} \triangleq \text{vec}(\mathbf{U}) \in \mathbb{R}^{rd_1}$. Denote by $\mathbf{P}_{\mathbf{V}} \in \mathbb{R}^{n\times n}$ the projection operator onto the column space of

$\mathbf{X}_{\mathbf{V}}$. Therefore, we may further bound the above display by

$$\begin{aligned}
\Big(\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \mathbf{\Theta}^*\rangle_{\text{HS}}^2\Big)^{1/2} &\leq \left(\frac{4}{n}\right)^{1/2} \sup_{\substack{\mathbf{U}\in\mathbb{R}^{d_1\times 2r}\\ \mathbf{V}\in\mathbb{R}^{d_2\times 2r}\\ \sum_{i=1}^{n}\langle \mathbf{X}_i\mathbf{V},\mathbf{U}\rangle_{\text{HS}}^2>0}} \frac{\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i\mathbf{V}, \mathbf{U}\rangle_{\text{HS}}}{(\sum_{i=1}^{n}\langle \mathbf{X}_i\mathbf{V}, \mathbf{U}\rangle_{\text{HS}}^2)^{1/2}} \\
&\leq \left(\frac{4}{n}\right)^{1/2} \sup_{\substack{\mathbf{U}\in\mathbb{R}^{d_1\times 2r}\\ \mathbf{V}\in\mathbb{R}^{d_2\times 2r}\\ \|\mathbf{X}_{\mathbf{V}}\gamma_{\mathbf{U}}\|_2>0}} \frac{\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{X}_{\mathbf{V}}\gamma_{\mathbf{U}}}{\|\mathbf{X}_{\mathbf{V}}\gamma_{\mathbf{U}}\|_2} \\
&\leq \left(\frac{4}{n}\right)^{1/2} \sup_{\mathbf{V}\in\mathbb{R}^{d_2\times 2r}} \|\mathbf{P}_{\mathbf{V}}\boldsymbol{\varepsilon}\|_2,
\end{aligned}$$

where the last line follows from the Cauchy-Schwarz inequality and the identity $\mathbf{P}_{\mathbf{V}}\mathbf{X}_{\mathbf{V}} = \mathbf{X}_{\mathbf{V}}$. The conclusion follows immediately by squaring both sides. □

Proof of Theorem 5.3. Now, for a fixed $\mathbf{V} \in \mathbb{R}^{d_2\times 2r}$, there exists a matrix $\tilde{\mathbf{V}} \in \mathcal{N}_\delta(\mathcal{P})$ such that $\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\tilde{\mathbf{V}}}\|_{\text{HS}} \leq \delta$. Then,

$$\|\mathbf{P}_{\mathbf{V}}\boldsymbol{\varepsilon}\|_2^2 \leq 2\|(\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\tilde{\mathbf{V}}})\boldsymbol{\varepsilon}\|_2^2 + 2\|\mathbf{P}_{\tilde{\mathbf{V}}}\boldsymbol{\varepsilon}\|_2^2 \leq 2\delta^2\|\boldsymbol{\varepsilon}\|_2^2 + 2\|\mathbf{P}_{\tilde{\mathbf{V}}}\boldsymbol{\varepsilon}\|_2^2.$$

Define $\mathcal{T} = \mathcal{T}_n$ as

$$\mathcal{T} \triangleq \bigcap_{\tilde{\mathbf{V}}\in\mathcal{N}_\delta(\mathcal{P})} \{\|\mathbf{P}_{\tilde{\mathbf{V}}}\boldsymbol{\varepsilon}\|_2^2 \leq a_2 r\max(d_1, d_2\log(d_1n^3/\delta)) + 2\sigma_\varepsilon^2 r d_2\} \cap \{\|\boldsymbol{\varepsilon}\|_2^2 \leq a_1 n + n\sigma_\varepsilon^2\}$$

for some constants $a_1, a_2 \geq \max(1, K_\varepsilon^2)$ to be chosen later. By the Hanson-Wright inequality (Theorem 1.1 of Rudelson and Vershynin (2013)), it follows that

$$\begin{aligned}
&\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2^2 > t + \sigma_\varepsilon^2 n) \leq 2\exp\left[-a_3\min\left(\frac{t^2}{nK_\varepsilon^4}, \frac{t}{K_\varepsilon^2}\right)\right], \\
&\mathbb{P}(\|\mathbf{P}_{\tilde{\mathbf{V}}}\boldsymbol{\varepsilon}\|_2^2 > t + 2\sigma_\varepsilon^2 r d_2) \leq 2\exp\left[-a_3\min\left(\frac{t^2}{2rd_2K_\varepsilon^4}, \frac{t}{K_\varepsilon^2}\right)\right],
\end{aligned}$$

for some universal constant $a_3 > 0$. Hence, a union bound implies

$$\begin{aligned}
\mathbb{P}(\mathcal{T}^{\mathsf{c}}) &\leq 2\exp\left[-a_2a_3K_\varepsilon^{-2}r\max(d_1, d_2\log(d_1n^3/\delta)) + N_\delta(\mathcal{P})\right] + 2\exp[-a_1a_3K_\varepsilon^{-2}n] \\
&\leq 2\exp\left[-a_2a_3K_\varepsilon^{-2}r\max(d_1, d_2\log(d_1n^3/\delta)) + 2rd_1\log(2) + (2rd_2+1)\log(24rd_1n^3/\delta)\right] \\
&\quad + 2\exp[-a_1a_3K_\varepsilon^{-2}n] \\
&\leq 2\exp[-a_4r\max(d_1, d_2\log(d_1n^3/\delta))] + 2\exp[-a_4n].
\end{aligned}$$

for some constant $a_4 > 0$ depending on K_ε, a_1, a_2, and a_3. Now, on the event $\mathcal{T}$, it follows that

$$\|\mathbf{P}_{\mathbf{V}}\boldsymbol{\varepsilon}\|_2^2 \leq 2\delta^2(a_1 + \sigma_\varepsilon^2)n + 2a_2 r \max(d_1, d_2 \log(d_1 n^3/\delta)) + 4rd_2\sigma_\varepsilon^2.$$

Letting $\delta = rn^{-1}\max(d_1, d_2)$, we have

$$\|\mathbf{P}_{\mathbf{V}}\boldsymbol{\varepsilon}\|_2^2 \leq 2(a_1 + \sigma_\varepsilon^2)r^2 n^{-1}\max(d_1^2, d_2^2) + 2a_2 r \max(d_1, d_2 \log(r^{-1}n^2 \min(1, d_1 d_2^{-1}))) + 4rd_2\sigma_\varepsilon^2.$$

Since this holds for an arbitrary $\mathbf{V} \in \mathbb{R}^{d_2 \times 2r}$, we conclude that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0} - \boldsymbol{\Theta}^*\rangle_{\text{HS}}^2 \leq 8(a_1 + \sigma_\varepsilon^2)r^2\max(d_1^2, d_2^2)n^{-2} + 16\sigma_\varepsilon^2 r d_2 n^{-1} \\ + 8a_2 r \max(d_1, d_2\log(r^{-1}n^2\min(1, d_1 d_2^{-1})))n^{-1}.\end{aligned}$$

Let $c_1 = 8a_1 + 24\sigma_\varepsilon^2 + 8a_2$ and $c_2 = a_4$. Using the fact that $d_1 \leq d_2 = d$ and $rd < n$ finishes the proof. □

Proof of Corollary 5.3.1. Let $\tilde{\boldsymbol{\Theta}}$ be defined as

$$\tilde{\boldsymbol{\Theta}} \triangleq \tilde{\boldsymbol{\Theta}}(r) = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}, \text{rank}(\boldsymbol{\Theta})\leq r}{\arg\min} \sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \boldsymbol{\Theta}\rangle_{\text{HS}})^2.$$

Then, by the definition of $\hat{\boldsymbol{\Theta}}_{L_0}$, we have that

$$\sum_{i=1}^{n}(y_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2 \leq \sum_{i=1}^{n}(y_i - \langle \mathbf{X}_i, \tilde{\boldsymbol{\Theta}}\rangle_{\text{HS}})^2.$$

Expanding the square and rearranging yields

$$\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2 \leq \sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\boldsymbol{\Theta}}\rangle_{\text{HS}})^2 + 2\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0} - \tilde{\boldsymbol{\Theta}}\rangle_{\text{HS}}.$$

If $\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2 = 0$, then the result immediately follows. Hence, for the remainder of the proof, we assume that $\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2 > 0$. Now, dividing both sides, it follows that

$$\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2\Big\}^{1/2} \leq \frac{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\boldsymbol{\Theta}}\rangle_{\text{HS}})^2}{\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2\Big\}^{1/2}} + 2\frac{\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0} - \tilde{\boldsymbol{\Theta}}\rangle_{\text{HS}}}{\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\boldsymbol{\Theta}}_{L_0}\rangle_{\text{HS}})^2\Big\}^{1/2}}.$$

By the construction of $\tilde{\mathbf{\Theta}}$, we deduce that

$$\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}})^2 \le \sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2$$

and

$$\begin{aligned}\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}}^2 &\le 2\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2 + 2\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}})^2 \\ &\le 4\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2.\end{aligned}$$

Therefore, we have that

$$\frac{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}})^2}{\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2\Big\}^{1/2}} \le \Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}})^2\Big\}^{1/2}.$$

Moreover, note that $\mathrm{rank}(\hat{\mathbf{\Theta}}_{L_0} - \tilde{\mathbf{\Theta}}) \le 2r$; hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{i=1}^{n}\varepsilon_i\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}} &\le \sup_{\mathbf{V}\in\mathbb{R}^{d_2\times 2r}}\|\mathbf{P}_{\mathbf{V}}\|_2\Big\{\sum_{i=1}^{n}\langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} - \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}}^2\Big\}^{1/2} \\ &\le 4\sup_{\mathbf{V}\in\mathbb{R}^{d_2\times 2r}}\|\mathbf{P}_{\mathbf{V}}\|_2\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2\Big\}^{1/2}.\end{aligned}$$

Combining these calculations, it follows that

$$\Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}\rangle_{\mathrm{HS}})^2\Big\}^{1/2} \le \Big\{\sum_{i=1}^{n}(f_i - \langle \mathbf{X}_i, \tilde{\mathbf{\Theta}}\rangle_{\mathrm{HS}})^2\Big\}^{1/2} + 8\sup_{\mathbf{V}\in\mathbb{R}^{d_2\times 2r}}\|\mathbf{P}_{\mathbf{V}}\|_2.$$

It is left to bound $\sup_{\mathbf{V}\in\mathbb{R}^{d_2\times 2r}}\|\mathbf{P}_{\mathbf{V}}\|_2^2$, which is provided in the proof of Theorem 5.3. □

Proof of Lemma 5.4. Temporarily fix $\pi \in \tilde{\Pi}$. Since $f^{(\pi)}(x_i) = 0$ for $i \in \mathcal{A}_1^{(\pi)} \cup \mathcal{A}_2^{(\pi)}$ by construction, we have that

$$\sum_{i=1}^{n}[f^{(\pi)}(x_i)]^2 = \sum_{i\in\mathcal{A}_3^{(\pi)}}[f^{(\pi)}(x_i)]^2 \le \sum_{i\in\mathcal{A}_3^{(\pi)}} f^2(x_{\pi(i)}) \le 2\log^2(n)\sigma_f^2 + c_1\log(n)$$

for some constant $c_1 > 0$ by Chebyshev's inequality when n is sufficiently large with probability at least $1-\delta$. Since $\pi \in \tilde{\Pi}$ is arbitrary, this finishes the proof. □

Proof of Theorem 5.5. The proof is standard. For example, see Section 15.2 of Lehmann and Romano (2006). □

Proof of Theorem 5.6. Fix $0 < \delta < \alpha(1-\alpha)/4$. Then, by the triangle inequality, it follows that

$$\Lambda^{(\pi_0)} = \sum_{i=1}^{n}[\hat{f}^{(\pi_0)}(x_i)]^2 \geq 2^{-1}\sum_{i=1}^{n}[f^{(\pi_0)}(x_i)]^2 - \sum_{i=1}^{n}[\hat{f}^{(\pi_0)}(x_i) - f^{(\pi_0)}(x_i)]^2.$$

From Assumption (5.2), the Chebyshev's inequality implies that there exists a constant $t_1 > 0$ (not depending on n) such that, for n sufficiently large,

$$\sum_{i=1}^{n}[f^{(\pi_0)}(x_i)]^2 \geq n\sigma_f^2 - t_1 n^{1/2} \tag{A.4.6.1}$$

with probability at least $1-\delta$. Assumption (5.3) ensures that

$$\sum_{i=1}^{n}[\hat{f}^{(\pi_0)}(x_i) - f^{(\pi_0)}(x_i)]^2 \leq \ell_n$$

with probability at least $1-\delta$. Hence, it holds with probability at least $1-2\delta$ that

$$\Lambda^{(\pi_0)} \geq 2^{-1}(n\sigma_f^2 - t_1 n^{1/2}) - \ell_n.$$

Now, temporarily fix $\pi \in \tilde{\Pi}$. Again, by the triangle inequality, it follows that

$$\Lambda^{(\pi)} \leq 2\sum_{i=1}^{n}[\hat{f}^{(\pi)}(x_i)]^2 - f^{(\pi)}(x_i)]^2 + 2\sum_{i=1}^{n}[f^{(\pi)}(x_i)]^2$$

Assumption (5.3) implies that

$$\sum_{i=1}^{n}[\hat{f}^{(\pi)}(x_i)]^2 - f^{(\pi)}(x_i)]^2 \leq \ell_n$$

with probability at least $1-\delta$. Moreover, we have from Lemma 5.4 that

$$\sum_{i\in\mathcal{A}_3^{(\pi)}}[f^{(\pi)}(x_i)]^2 \leq \log^2(n)\sigma_f^2 + t_2\log(n)$$

with probability at least $1-\delta$ for some constant $t_2 > 0$. Hence, with probability at least $1-2\delta$,

$$\Lambda^{(\pi)} \leq 2\ell_n + 2\log^2(n)\sigma_f^2 + 2t_2\log(n).$$

Combining the above calculations, for n sufficiently large,

$$\begin{aligned}\Lambda^{(\pi_0)} - \Lambda^{(\pi)} &\geq 2^{-1}(n\sigma_f^2 - t_1 n^{1/2}) - 3\ell_n - 2\log^2(n)\sigma_f^2 - 2t_2\log(n) \\ &\geq 2^{-1}\{h(n^{1/2} + \ell_n) - t_1 n^{1/2}\} - 3\ell_n - 2h(n^{-1/2} + \ell_n n^{-1})\log^2(n) - 2t_2\log(n) \\ &> 0\end{aligned}$$

with probability at least $1 - 4\delta$ if $h > 0$ is sufficiently large (not depending on n). Thus,

$$\mathbb{P}_{H_1}(\Lambda^{(\pi_0)} > \Lambda^{(\pi)}) \geq 1 - 4\delta$$

for n sufficiently large. Since π is arbitrary, it follows that

$$\liminf_{n\to\infty} \min_{\pi\in\tilde{\Pi}} \mathbb{P}_{H_1}(\Lambda^{(\pi_0)} > \Lambda^{(\pi)}) \geq 1 - 4\delta.$$

Hence,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}_{H_1}\varphi &= \limsup_{n\to\infty} |\Pi|^{-1}\mathbb{E}_{H_1}\sum_{\pi\in\Pi} \mathbb{1}_{\Lambda^{(\pi_0)}\leq\Lambda^{(\pi)}} \\ &= 1 - \liminf_{n\to\infty} |\Pi|^{-1}\sum_{\pi\in\tilde{\Pi}} \mathbb{P}_{H_1}(\Lambda^{(\pi_0)} > \Lambda^{(\pi)}) \\ &\leq 4\delta.\end{aligned}$$

Since $\delta < \alpha(1-\alpha)/4$, the result follows from Markov's inequality. $\square$

Proof of Corollary 5.6.1. By Chebyshev's inequality, there exists a constant $t_3 > 0$ such that

$$\sum_{i=1}^n [f^{(\pi_0)}(x_i)]^2 \geq n\sigma_f^2 - t_3 n^{1/2}\sigma_f^2$$

with probability at least $1 - \delta$. The remainder of the proof is identical, replacing the bound in equation (A.4.6.1) with the above bound. $\square$

Proof of Theorem 5.7. It is immediate from the definition of $\hat{f}_{\mathrm{LA}}(\cdot)$ that $\hat{f}_{\mathrm{LA}}(\mathbf{x}_i; (\mathbf{x}_j, y_j)_{j=1}^n; \pi) = \hat{f}_{\mathrm{LA}}(\mathbf{x}_i; (\mathbf{x}_j, y_{\pi(j)})_{j=1}^n; \pi_0)$ for any $\pi \in \Pi$. Now, if $\pi = \pi_0$, the compatibility condition for the design is satisfied for some constant φ_{cc} with probability at least $1 - \delta/2$. Then, it follows from Theorem 6.1 of Bühlmann and van de Geer (2011) that

$$\mathbb{P}\Big\{\sum_{i=1}^n \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{LA}} - \boldsymbol{\beta}^*\rangle_2^2 \leq c_2\lambda^2 sn/\varphi_{\mathrm{cc}}^2\Big\} \geq 1 - \delta/2$$

for some constant $c_2 > 0$. Now, let $\pi \in \tilde{\Pi}$ be arbitrary and define the event

$$\mathcal{T} \triangleq \Big\{ \max_{j \in [p]} \Big| \sum_{i=1}^{n} x_{i,j} y_{\pi(i)} \Big| \leq 3 c_3^{-1/2} n^{1/2} L_\xi \log^{1/2}(6p/\delta) \Big\},$$

where $x_{i,j}$ denotes the jth entry of $\mathbf{x}_i$. Fix $j \in [p]$ and let $\xi_{i,j} \triangleq x_{i,j} y_{\pi(i)}$. By the triangle inequality, we have that

$$\Big| \sum_{i=1}^{n} \xi_{i,j} \Big| \leq \Big| \sum_{i \in \mathcal{A}_1^{(\pi)}} \xi_{i,j} \Big| + \Big| \sum_{i \in \mathcal{A}_2^{(\pi)}} \xi_{i,j} \Big| + \Big| \sum_{i \in \mathcal{A}_3^{(\pi)}} \xi_{i,j} \Big|.$$

Then, by the construction of $\mathcal{A}_1^{(\pi)}$, we have that $(\xi_{i,j})_{i \in \mathcal{A}_1^{(\pi)}}$ are independent and identically distributed sub-exponential random variables with parameter $L_\xi \leq K_x(K_f + K_\varepsilon)$. By Bernstein's inequality, for any $t_1 > 0$, it follows that

$$\mathbb{P}\Big(\Big| \sum_{i \in \mathcal{A}_1^{(\pi)}} \xi_{i,j} \Big| > t_1 \Big) \leq 2 \exp \Big[- c_3 \min(|\mathcal{A}_1^{(\pi)}|^{-1} t_1^2 L_\xi^{-2}, t_1 L_\xi^{-1}) \Big]$$

for some universal constant $c_3 > 0$. Let

$$t_1 \triangleq c_1^{-1/2} n^{1/2} L_\xi \log^{1/2}(6p/\delta).$$

Noting that $|\mathcal{A}_1^{(\pi)}| \leq n$, we have for n sufficiently large,

$$\mathbb{P}\Big(\Big| \sum_{i \in \mathcal{A}_1^{(\pi)}} \xi_{i,j} \Big| > c_3^{-1/2} n^{1/2} L_\xi \log^{1/2}(6p/\delta) \Big) \leq \delta/(3p).$$

Taking a union bound shows that

$$\mathbb{P}\Big(\max_{j \in [p]} \Big| \sum_{i \in \mathcal{A}_1^{(\pi)}} \xi_{i,j} \Big| > c_3^{-1/2} n^{1/2} L_\xi \log^{1/2}(6p/\delta) \Big) \leq \delta/3.$$

A similar calculation for $\mathcal{A}_2^{(\pi)}$ yields

$$\mathbb{P}\Big(\max_{j \in [p]} \Big| \sum_{i \in \mathcal{A}_2^{(\pi)}} \xi_{i} \Big| > c_3^{-1/2} n^{1/2} L_\xi \log^{1/2}(6p/\delta) \Big) \leq \delta/3.$$

Now,

$$\Big|\sum_{i\in\mathcal{A}_3^{(\pi)}}\xi_{i,j}\Big| \le |\mathcal{A}_3^{(\pi)}|\max_{j\in[p]}\max_{i\in\mathcal{A}_3^{(\pi)}}|\xi_{i,j}|.$$

Again, by Bernstein's inequality, for n sufficiently large,

$$\mathbb{P}\Big\{\max_{j\in[p]}\max_{i\in\mathcal{A}_3^{(\pi)}}|\xi_{i,j}| > c_3^{-1}L_\xi\log(6p|\mathcal{A}_3^{(\pi)}|/\delta)\Big\} \le \delta/3.$$

Combining the above calculations, we have that

$$\mathbb{P}(\mathcal{T}) \ge 1-\delta$$

for n sufficiently large. On $\mathcal{T}$, for any $\boldsymbol{\beta}\in\mathbb{R}^p$, it follows that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_{\pi(i)}-\langle\mathbf{x}_i,\boldsymbol{\beta}\rangle_2)^2+\lambda\|\boldsymbol{\beta}\|_1 &= \frac{1}{n}\{\|\mathbf{y}^{(\pi)}\|_2^2-2\langle\mathbf{y}^{(\pi)},\mathbf{X}\boldsymbol{\beta}\rangle_2+\|\mathbf{X}\boldsymbol{\beta}\|_2^2\}+\lambda\|\boldsymbol{\beta}\|_1\\
&\ge \frac{1}{n}\{\|\mathbf{y}^{(\pi)}\|_2^2-6c_3^{-1/2}n^{1/2}L_\xi\log^{1/2}(6p/\delta)\|\boldsymbol{\beta}\|_1+\|\mathbf{X}\boldsymbol{\beta}\|_2^2\}+\lambda\|\boldsymbol{\beta}\|_1\\
&\ge \frac{1}{n}\{\|\mathbf{y}^{(\pi)}\|_2^2+\|\mathbf{X}\boldsymbol{\beta}\|_2^2\}+(\lambda-2\lambda_0)\|\boldsymbol{\beta}\|_1.
\end{aligned}$$

Thus, the above is minimized when $\boldsymbol{\beta}=\mathbf{0}_p$. Therefore,

$$\mathbb{P}(\hat{\boldsymbol{\beta}}_{\mathrm{LA}}^{(\pi)}=\mathbf{0}_p)\ge 1-\delta.$$

Invoking Corollary 5.3.1 finishes the proof. □

To facilitate the proof of Theorem 5.8, we define three auxiliary estimators. Let

$$\hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \triangleq \mathop{\arg\min}_{\boldsymbol{\beta}\in\mathbb{R}^p,\|\boldsymbol{\beta}\|=s}\sum_{i=\mathcal{A}_k^{(\pi)}}(y_{\pi(i)}-\langle\mathbf{x}_i,\boldsymbol{\beta}\rangle_2)^2$$

for $k=1,2,3$. Thus, $\hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})}$ is the L_0 estimator of $\boldsymbol{\beta}$ using only the data $(\mathbf{x}_i,y_{\pi(i)})_{i\in\mathcal{A}_k^{(\pi)}}$ for $k=1,2,3$. The following lemma relates the squared predicted values of $\hat{\boldsymbol{\beta}}_{L_0}^{(\pi)}$ with $(\hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})})_{k=1}^3$. The result allows us to decouple the dependence between the covariates and the response by analyzing the observations in $\mathcal{A}_1^{(\pi)}$ and $\mathcal{A}_2^{(\pi)}$ separately.

Lemma 5.11. *Consider the model given in equation* (A.3.1.1). *Then,*

$$\sum_{i=1}^{n} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0} \rangle_2^2 \leq \sum_{i \in \mathcal{A}_1^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_1^{(\pi)})} \rangle_2^2 + \sum_{i \in \mathcal{A}_2^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_2^{(\pi)})} \rangle_2^2 + \sum_{i \in \mathcal{A}_3^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_3^{(\pi)})} \rangle_2^2.$$

Proof of Lemma 5.11. Indeed, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have that

$$\sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_2)^2 = \sum_{k=1}^{3} \sum_{i \in \mathcal{A}_k^{(\pi)}} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_2)^2.$$

Minimizing both sides with respect to $\boldsymbol{\beta}$, it follows that

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\pi)} \rangle_2)^2 &= \min_{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 = s} \sum_{k=1}^{3} \sum_{i \in \mathcal{A}_k^{(\pi)}} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_2)^2 \\ &\geq \sum_{k=1}^{3} \min_{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 = s} \sum_{i \in \mathcal{A}_k^{(\pi)}} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_2)^2 \\ &= \sum_{k=1}^{3} \sum_{i \in \mathcal{A}_k^{(\pi)}} (y_i - \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \rangle_2)^2. \end{aligned}$$

Applying the Pythagorean Theorem finishes the proof. □

Proof of Theorem 5.8. It is clear that $\hat{f}_{L_0}(\mathbf{x}_i; (\mathbf{x}_j, y_j)_{j=1}^n; \pi) = \hat{f}_{L_0}(\mathbf{x}_i; (\mathbf{x}_j, y_{\pi(j)})_{j=1}^n; \pi_0)$ for any $\pi \in \Pi$. Moreover, from Theorem 2.6 of Rigollet and Hütter (2017), there exists a constant $c_1 > 0$ such that

$$\mathbb{P}\Big\{ \sum_{i=1}^{n} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0} - \boldsymbol{\beta}^* \rangle_2^2 \leq c_1 K_\varepsilon (\log \binom{p}{2s} + \log(1/\delta)) \Big\} \geq 1 - \delta.$$

Now, fix $\pi \in \tilde{\Pi}$. Since $(\mathbf{x}_i)_{i \in \mathcal{A}_1^{(\pi)}}$ and $(y_{\pi(i)})_{i \in \mathcal{A}_1^{(\pi)}}$ are mutually independent, Theorem 2.6 of Rigollet and Hütter (2017) implies there exists a constant $c_2 > 0$ such that

$$\mathbb{P}\Big\{ \sum_{i \in \mathcal{A}_1^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_1^{(\pi)})} \rangle_2^2 \leq c_2 (K_f + K_\varepsilon)(\log \binom{p}{2s} + \log(3/\delta)) \Big\} \geq 1 - \delta/3.$$

Analogously, we see that

$$\mathbb{P}\Big\{ \sum_{i \in \mathcal{A}_2^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_2^{(\pi)})} \rangle_2^2 \leq c_2(K_f + K_\varepsilon)(\log \binom{p}{2s} + \log(3/\delta)) \Big\} \geq 1 - \delta/3.$$

Next, by an argument identical to that of Lemma 5.4, it follows that

$$\sum_{i \in \mathcal{A}_1^{(\pi)}} \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0}^{(\mathcal{A}_3^{(\pi)})} \rangle_2^2 \leq \sum_{i \in \mathcal{A}_1^{(\pi)}} y_{\pi(i)}^2 \leq \log^2(n)(\sigma_f^2 + \sigma_\varepsilon^2) + c_3 \log(n)$$

with probability at least $1 - \delta/3$ for some constant $c_3 > 0$. Hence, Lemma 5.11 implies that

$$\sum_{i=1}^{n} (\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{L_0} \rangle_2)^2 \leq 2c_2(K_f + K_\varepsilon)(\log \binom{p}{2s} + \log(3/\delta)) + \log^2(n)(\sigma_f^2 + \sigma_\varepsilon^2) + c_3 \log(n)$$

with probability at least $1 - \delta$. The result now follows from Corollary 5.3.1. □

Proof of Theorem 5.9. The proof is identical to that of Theorem 5.8, replacing Theorem 2.6 of Rigollet and Hütter (2017) with Theorem 5.3 of the present paper. □

Proof of Theorem 5.10. Let $\tilde{\boldsymbol{\Theta}} = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^\top$ with $\tilde{\mathbf{U}} \in \mathbb{R}^{d_1 \times r}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{d_2 \times r}$, $\tilde{f}_i \triangleq \langle \mathbf{X}_i, \tilde{\boldsymbol{\Theta}} \rangle_{\text{HS}}$ with variance $\sigma_{\tilde{f}}^2 \triangleq \text{Var}(\tilde{f}_1) = \text{vec}(\tilde{\boldsymbol{\Theta}})^\top \boldsymbol{\Sigma} \text{vec}(\tilde{\boldsymbol{\Theta}})$, and $\eta_i \triangleq \langle \mathbf{X}_i, \boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}} \rangle_{\text{HS}}$, yielding the decomposition

$$y_i = \langle \mathbf{X}_i, \tilde{\boldsymbol{\Theta}} \rangle_{\text{HS}} + \eta_i + \varepsilon_i = \langle \mathbf{X}_i \tilde{\mathbf{V}}, \tilde{\mathbf{U}} \rangle_{\text{HS}} + \eta_i + \varepsilon_i.$$

Since $\tilde{\boldsymbol{\Theta}}$ satisfies

$$\tilde{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \text{rank}(\boldsymbol{\Theta}) \leq r}{\arg\min} \mathbb{E}\langle \mathbf{X}, \boldsymbol{\Theta}^* - \boldsymbol{\Theta} \rangle_{\text{HS}}^2,$$

it follows from the population first-order condition that

$$\mathbb{E}\text{vec}(\mathbf{X}\tilde{\mathbf{V}}) \langle \mathbf{X}, \boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}} \rangle_{\text{HS}} = \mathbb{E}\text{vec}(\mathbf{X}\tilde{\mathbf{V}})\eta = \mathbf{0}_{rd_1},$$

implying that $\text{vec}(\mathbf{X}_i \tilde{\mathbf{V}})$ is uncorrelated with η_i. Now, consider an auxiliary oracle estimator $\hat{\boldsymbol{\Theta}}_{\tilde{\mathbf{V}}}$ given by

$$\hat{\mathbf{U}}_{\tilde{\mathbf{V}}} \triangleq \underset{\mathbf{U} \in \mathbb{R}^{d_1 \times r}}{\arg\min} \sum_{i=1}^{n} (y_i - \langle \mathbf{X}_i, \mathbf{U}\tilde{\mathbf{V}}^\top \rangle_{\text{HS}})^2 \text{ and } \hat{\boldsymbol{\Theta}}_{\tilde{\mathbf{V}}} \triangleq \hat{\mathbf{U}}_{\tilde{\mathbf{V}}} \tilde{\mathbf{V}}^\top.$$

Since $\hat{\mathbf{\Theta}}_{L_0}$ is the empirical risk minimizer, it follows from the Pythagorean Theorem that

$$\Lambda^{(\pi_0)} = \sum_{i=1}^{n} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0} \rangle_{\text{HS}}^2 \geq \sum_{i=1}^{n} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{\tilde{\mathbf{V}}} \rangle_{\text{HS}}^2 = \|\mathbf{P}_{\tilde{\mathbf{V}}}\mathbf{y}\|_2^2 = \|\tilde{\boldsymbol{f}}\|_2^2 + 2\langle \tilde{\boldsymbol{f}}, \boldsymbol{\eta} + \boldsymbol{\varepsilon} \rangle_2 + \|\mathbf{P}_{\tilde{\mathbf{V}}}(\boldsymbol{\eta} + \boldsymbol{\varepsilon})\|_2^2.$$

Fix $\delta > 0$ arbitrarily. Now, proceeding as in the proof of Theorem 5.6 and Corollary 5.6.1, there exists $t_1 > 0$, depending on δ, $\|\tilde{\mathbf{\Theta}}\|_{\text{F}}$ and K_x, such that with probability at least $1 - \delta/2$,

$$\|\tilde{\boldsymbol{f}}\|_2^2 \geq n\sigma_{\tilde{f}}^2 - t_1 n^{1/2}.$$

If in addition Assumption (5.5*) is satisfied, then

$$\|\tilde{\boldsymbol{f}}\|_2^2 \geq n\sigma_{\tilde{f}}^2 - t_1' n^{1/2}\sigma_{\tilde{f}}^2$$

with probability at least $1-\delta/2$, where t_1' depends on δ and ϑ. For the second term, since $\text{vec}(\mathbf{X}_i\tilde{\mathbf{V}})$ is uncorrelated with η_i, it follows that $\mathbb{E}\langle \tilde{\boldsymbol{f}}, \boldsymbol{\eta} + \boldsymbol{\varepsilon} \rangle_2 = 0$. Now, by Chebyshev's inequality, there exists $t_2 > 0$ depending only on δ such that

$$2|\langle \tilde{\boldsymbol{f}}, \boldsymbol{\eta} + \boldsymbol{\varepsilon} \rangle_2| \leq t_2 n^{1/2} \sigma_{\tilde{f}} (\sigma_f^2 + \sigma_\varepsilon^2)$$

with probability at least $1 - \delta/2$. Therefore, with probability at least $1 - \delta$,

$$\Lambda^{(\pi_0)} \geq n\sigma_{\tilde{f}}^2 - t_1 n^{1/2} - t_2 n^{1/2} \sigma_{\tilde{f}} (\sigma_f^2 + \sigma_\varepsilon^2).$$

It remains to bound $\Lambda^{(\pi)}$ for $\pi \in \tilde{\Pi}$. Define the auxiliary estimators

$$\hat{\mathbf{\Theta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \triangleq \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \text{rank}(\mathbf{\Theta}) \leq r}{\arg\min} \sum_{i \in \mathcal{A}_k^{(\pi)}} (y_{\pi(i)} - \langle \mathbf{X}_i, \mathbf{\Theta} \rangle_{\text{HS}})^2.$$

By an identical argument as in Lemma 5.11, it follows that

$$\Lambda^{(\pi)} = \sum_{i=1}^{n} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}^{(\pi)} \rangle_{\text{HS}}^2 \leq \sum_{k=1}^{3} \sum_{i \in \mathcal{A}_k^{(\pi)}} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \rangle_{\text{HS}}^2.$$

By Theorem 5.3, there exists a constant t_3, depending on δ, σ_f^2, σ_ε^2, K_x, and K_ε, such that

$$\sum_{i \in \mathcal{A}_k^{(\pi)}} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \rangle_{\text{HS}}^2 \leq t_3 r d \log(n)$$

for $j = 1, 2$ with probability at least $1 - \delta/3$. Similarly, following Lemma 5.4, we have that

$$\sum_{i \in \mathcal{A}_k^{(\pi)}} \langle \mathbf{X}_i, \hat{\mathbf{\Theta}}_{L_0}^{(\mathcal{A}_k^{(\pi)})} \rangle_{\text{HS}}^2 \leq \log^2(n)(\sigma_f^2 + \sigma_\varepsilon^2) + t_4 \log(n)$$

with probability at least $1 - \delta/3$ for a constant t_4 depending on δ, $\|\mathbf{\Theta}^*\|_{\mathrm{F}}$, K_x, and K_ε. Combining, we have

$$\Lambda^{(\pi)} \leq 2t_3 rd \log(n) + \log^2(n)(\sigma_f^2 + \sigma_\varepsilon^2) + t_4 \log(n)$$

with probability at least $1 - \delta$. Proceeding as in Theorem 5.6 finishes the proof. $\square$

BIBLIOGRAPHY

Achieser, N. I. (1992). *Theory of approximation*. Dover Publications.

Bai, R., Boland, M. R., and Chen, Y. (2019). Fast algorithms and theory for high-dimensional bayesian varying coefficient models. *arXiv preprint arXiv:1907.06477*.

Basu, S., Pollack, R., and Coste-Roy, M. (2007). *Algorithms in Real Algebraic Geometry*. Algorithms and Computation in Mathematics. Springer Berlin Heidelberg.

Bellec, P. C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press.

Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2):721–741.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Boland, P. J., Singh, H., and Cukic, B. (2002). Stochastic orders in partition and random testing of software. *Journal of applied probability*, 39(3):555–565.

Bradic, J., Claeskens, G., and Gueuning, T. (2019). Fixed effects testing in high-dimensional linear mixed models. *Journal of the American Statistical Association*, (just-accepted):1–35.

Brown, L. D., Mukherjee, G., and Weinstein, A. (2018). Empirical bayes estimates for a two-way cross-classified model. *The Annals of Statistics*, 46(4):1693–1720.

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

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, 45(2):615–646.

Cai, T. T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):391–419.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5):2330–2355.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Candès, E. J. and Plan, Y. (2011). 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.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Chao, M.-T. and Strawderman, W. (1972). Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431.

Chen, F., Li, Z., Shi, L., and Zhu, L. (2015). Inference for mixed models of anova type with high-dimensional data. *Journal of Multivariate Analysis*, 133:382–401.

Chen, X. and He, Y. (2018). Inference of high-dimensional linear models with time-varying coefficients. *Statistica Sinica*, pages 255–276.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018b). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Efron, B. (2019). Bayes, oracle bayes and empirical bayes. *Statistical science*, 34(2):177–201.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J., Wang, W., and Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239.

Finucane, M. M., Paciorek, C. J., Danaei, G., and Ezzati, M. (2014). Bayesian estimation of population-level trends in measures of health status. *Statistical Science*, pages 18–25.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Godsil, C. and Royle, G. F. (2013). *Algebraic graph theory*, volume 207. Springer Science & Business Media.

Greenshtein, E. and Ritov, Y. (2019). Comment: Empirical bayes, compound decisions and exchangeability. *Statistical science*, 34(2):224–228.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, 24(2):137–154.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Guo, Z., Renaux, C., Bühlmann, P., and Cai, T. (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics*, 15(2):6633–6676.

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, 114(525):358–369.

Hartley, H. O. and Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17.

Jackson, D. (1941). *Fourier series and orthogonal polynomials*. Mathematical Association of America.

Janson, L., Barber, R. F., and Candes, E. (2017). Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1037–1065.

Javanmard, A. and Lee, J. D. (2020). A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):685–718.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Jiang, J. (1996). Reml estimation: asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, 44(5):2127–2160.

Klopp, O. and Pensky, M. (2015). Sparse high-dimensional varying coefficient model: Nonasymptotic minimax study. *The Annals of Statistics*, 43(3):1273–1299.

Knafl, G., Sacks, J., and Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association*, 80(391):683–691.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Law, M. and Ritov, Y. (2021a). High-dimensional varying coefficient models with functional random effects. *arXiv preprint arXiv:2110.06426*.

Law, M. and Ritov, Y. (2021b). Inference without compatibility: Using exponential weighting for inference on a parameter of a linear model. *Bernoulli*, 27(3):1467–1495.

Law, M. and Ritov, Y. (2022). Inference and estimation for random effects in high-dimensional linear mixed models. *Journal of the American Statistical Association*, pages 1–10.

Law, M., Ritov, Y., Zhang, R., and Zhu, Z. (2021). Rank-constrained least-squares: Prediction and inference. *arXiv preprint arXiv:2111.14287*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410.

Li, S., Cai, T. T., and Li, H. (2019). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *arXiv preprint arXiv:1907.06116*.

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, pages 1563–1588.

Lu, Y. and Zhang, G. (2010). The equivalence between likelihood ratio test and f-test for testing variance component in a balanced one-way random effects model. *Journal of Statistical Computation and Simulation*, 80(4):443–450.

Martin, M. O., Mullis, I. V., and Hooper, M. (2016). Methods and procedures in timss 2015. *TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA)*.

Mathew, T. and Sinha, B. K. (1988). Optimum tests in unbalanced two-way models without interaction. *The Annals of Statistics*, 16(4):1727–1740.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Mullis, I., Martin, M., Foy, P., and Hooper, M. (2016a). Timss 2015 international results in mathematics. retrieved from boston college, timss & pirls international study center.

Mullis, I. V., Martin, M. O., and Loveless, T. (2016b). 20 years of timss: International trends in mathematics and science achievement, curriculum, and instruction. *TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA)*.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697.

Qeadan, F. and Christensen, R. (2020). On the equivalence between the lrt and f-test for testing variance components in a class of linear mixed models. *Metrika: International Journal for Theoretical and Applied Statistics*, pages 1–26.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67.

Rigollet, P. and Hütter, J.-C. (2017). High-dimensional statistics. Technical report, Massachusetts Institute of Technology.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771.

Rodriguez-Martinez, A., Zhou, B., Sophiea, M. K., Bentham, J., Paciorek, C. J., Iurilli, M. L., ..., and Ezzati, M. (2020). Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: a pooled analysis of 2181 population-based studies with 65 million participants. *The Lancet*, 396(10261):1511–1524.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Stanley, R. P. (2012). *Enumerative Combinatorics: Volume 1.* Number Vol. 49 in Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515.

Wu, C. O., Chiang, C.-T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93(444):1388–1402.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *The Journal of Machine Learning Research*, 13(1):1973–1998.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948.