EMPIRICAL LIKELIHOOD BASED FUNCTIONAL DATA ANALYSIS AND HIGH DIMENSIONAL INFERENCE WITH APPLICATIONS TO BIOLOGY

By

Honglang Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2015

ABSTRACT

EMPIRICAL LIKELIHOOD BASED FUNCTIONAL DATA ANALYSIS AND HIGH DIMENSIONAL INFERENCE WITH APPLICATIONS TO BIOLOGY

By

Honglang Wang

High dimensional data analysis has been a rapidly developing topic in statistics with various applications in areas such as genetics/genomics, neuroscience, finance, social science and so on. With the rapid development of technology, statistics as a data science requires more and more innovations in methodologies as well as breakthroughs in mathematical frameworks. In high dimensional world, classical statistical methods designed for fixed dimensional models are often doomed to fail. This thesis focuses on two types of high dimensional data analysis. One is the study of typical "large p small n" problem in linear regression with high dimensional covariates $\mathbf{X} \in \mathbb{R}^p$ but small sample size n, and the other is the functional data analysis. Functional data belong to the class of high dimensional data in the sense that every data object consists of a large number of measurements, which may be larger than the sample size. But the key characteristic is that functional objects can be modeled as smooth curves or surfaces. We make use of Empirical Likelihood (EL) introduced by [Owe01], to solve some fundamental problems in these two particular high dimensional problems.

The first part of the thesis considers the problem of testing functional constraints in a class of functional linear regression models where both the predictors and the response are functional data measured at discrete time points. We propose test procedures based on the empirical likelihood with bias-corrected estimating equations to conduct both pointwise and simultaneous inference. The asymptotic distributions of the test statistics are derived

under the null and local alternative hypotheses, where sparse and dense functional data are considered in a unified framework. We find a phase transition in the asymptotic distributions and the orders of detectable alternatives from sparse to dense functional data. Specifically, the proposed tests can detect alternatives of root-n order when the number of repeated measurements per curve is of an order larger than n^{η_0} with n being the number of curves. The transition points η_0 are different for pointwise and simultaneous tests and both are smaller than the transition point in the estimation problem.

In the second part of the thesis, we consider hypothesis testing problems for a low-dimensional coefficient vector in a high-dimensional linear model under heteroscedastic error. Heteroscedasticity is a commonly observed phenomenon in many applications including finance and genomic studies. Several statistical inference procedures have been proposed for low-dimensional coefficients in a high-dimensional linear model with homoscedastic noise. However, those procedures designed for homoscedastic error are not applicable for models with heteroscedastic error and the heterscedasticity issue has not been investigated and studied. We propose a inference procedure based on empirical likelihood to overcome the heteroscedasticity issue. The proposed method is able to make valid inference under heteroscedasticity model even when the conditional variance of random error is a function of the high-dimensional predictor. We apply our inference procedure to three recently proposed estimating equations and establish the asymptotic distributions of the proposed methods.

For both of the two parts, simulation studies and real data analyses are conducted to demonstrate the proposed methods.

I would like to dedicate this thesis to my beloved parents, Daofu Wang and Shuizhen Wang, and my little brother, Hailang Wang.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my two official advisors Dr. Yuehua Cui and Dr. Ping-Shou Zhong for their continuous support guidance, understanding, patience and encouragement during my PhD study and research. Dr. Cui and Dr. Zhong have pushed me into contact with a multitude of disciplines, and their guidance about how to approach research, write, and give talks has been invaluable. They have also provided me excellent environments for doing research in the development of methodology and theory as well as in the real data analysis. Without their guidance and persistent help, this dissertation would not have been possible. For all of this, I am very thankful to both of my advisors.

I would also like to thank the other wonderful members of my research committee, Dr. C. Robin Buell and Dr. Hyokyoung (Grace) Hong. In getting my dual PhD degree in Quantitative Biology, Dr. Buell has provided much guidance and assistance, which also make me more confident to become a Bio-statistician. I have been enjoying involvement in the potato project and learning a lot from monthly group calls, annual meetings and the Bioinformatics workshops. Her guidance provided me with the unique opportunity to gain a wider breadth of experience in biology science, which is especially important for a Bio-statistician.

Besides that, I thank all the other professors and staff in this wonderful Department of Statistics and Probability who have never flinched about answering a question from a nagging graduate student—something that is embedded in the culture of Wells Hall. My special thanks go to Dr. Hira L. Koul, Dr. Yimin Xiao and Dr. Tapabrata Maiti for their interesting courses, valuable advise and encouragement.

Coming to friends, I am grateful to Yuzhen Zhou, Tao He, Jikai Lei, Chen Yue, Xin Qi,

Liqian Cai, Xiaoqing Zhu, Bin Gao, Xu Liu and all other fellow students from the Department of Statistics and Probability for the friendship and the fun time we spent together in the past five years.

Finally and most importantly, I would express my profound gratitude to my beloved parents, Daofu Wang and Shuizhen Wang and my litter brother, Hailang Wang for their love, endless support and faith in me in all of my endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

- EL: Empirical Likelihood;
- IID: Independent and Identically Distributed;
- KL: Kullback-Leibler;
- SNPs: Single Nucleotide Polymorphisms;
- GWAS: Genome Wide Association Studies;
- KFC: Key conFounder Controlling
- WGCNA: Weighted Gene Co-expression Network Analysis;

Chapter 1

Introduction

1.1 Empirical Likelihood

In Statistics, the likelihood principle is the primary principle as stated in [Edw84],

> *Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information.*

However, for the inference procedure to be more widely applicable, some non-parametric version of the likelihood is desirable so that we can not only gain robustness and flexibility but also keep the effectiveness as well as some other merits of the likelihood principle. In the late eighties, Professor Art B. Owen proposed the great idea, "Empirical Likelihood" (EL) [Owe88, Owe90], which is a non-parametric likelihood. The well known "Wilks Phenomenon" belonging to the parametric likelihood still holds for EL [Owe90, Owe01]. EL also enjoys the Bartlett correction property [DHR91, CC06]. Besides, it produces more natural data driven shape of confidence regions.

We consider the univariate mean inference problem to introduce the EL idea. Given n IID observations $\{X_i \in \mathbb{R}, i = 1, 2, \cdots, n\}$ from an unknown underlying distribution F_0 with finite first two moments, we want to conduct the inference for the univariate mean $\mu_0 := \mathbb{E}_{F_0}(X_i)$. A natural point estimation of μ_0 is the sample mean $\bar{X}$, but how to get

an efficient confidence interval with a given confidence level is not that simple since we have no idea about the underlying distribution up to the first two finite moments. According to [Owe90], the empirical likelihood for μ is the product of the probability weights, say $\{0 \leq p_i \leq 1, i = 1, 2, \cdots, n\}$, sitting on the sample points $\{X_i, i = 1, 2, \cdots, n\}$, that is $\prod_{i=1}^{n} p_i$, with the first moment constraint $\sum_{i=1}^{n} p_i X_i = \mu$, i.e.

$$\mathcal{L}_{\mathrm{EL}}(\mu) = \max_{\{p_i\}_{i=1}^{n}} \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, p_i \geq 0, \sum_{i=1}^{n} p_i (X_i - \mu) = 0 \right\}. \tag{1.1.1}$$

Actually, we can derive the above formulation (1.1.1) in the following formal way. The statistical model with the first moment restriction could be phrased formally as the set of all probability measures that are compatible with the first moment condition, i.e. $\mathcal{P} = \bigcup_{\mu} \mathcal{P}(\mu)$, where

$$\mathcal{P}(\mu) = \left\{ \text{probability measure } P \text{ on } \mathbb{R} : \int (X - \mu) dP = 0 \right\}.$$

Note that it is correctly specified if and only if $\mathcal{P}$ includes the true measure $dF_0(x)$ as its member. The following function could be regarded as a measure for the divergence between two probability measures P and Q:

$$D(P, Q) = \int \phi(\frac{dP}{dQ}) dQ,$$

as long as ϕ is chosen to be convex. And we know that the Kullback-Leibler (KL) divergence between probability measures P and Q is a special case by taking $\phi(x) = -\log(x)$.

If the model is correctly specified, we have the following nice property at the population level

$$\mu_0 = \inf_{\mu} \inf_{P \in \mathcal{P}(\mu)} D(P, F_0).$$

Hence a natural statistical procedure for the estimation of the mean can be obtained by replacing the unknown F_0 with the empirical measure F_n and searching over the restricted statistical model $\mathfrak{P} = \bigcup_\mu \mathfrak{P}(\mu)$, where

$$\mathfrak{P}(\mu) = \left\{ F_p := \sum_{i=1}^n p_i \delta_{X_i} : \int (X - \mu) dF_p = 0 \right\}.$$

And then the estimation of the mean is defined as the minimizer of the following optimization problem

$$\inf_\mu \inf_{F_p \in \mathfrak{P}(\mu)} D(F_p, F_n) = \inf_\mu \inf_{\substack{\sum_{i=1}^n p_i (X_i - \mu) = 0, \\ \sum_{i=1}^n p_i = 1, p_i \geq 0}} \frac{1}{n} \sum_{i=1}^n \phi(n p_i). \tag{1.1.2}$$

In particular, with the KL divergence in (1.1.2), we have

$$\inf_\mu \inf_{\substack{\sum_{i=1}^n p_i (X_i - \mu) = 0, \\ \sum_{i=1}^n p_i = 1, p_i \geq 0}} \frac{1}{n} \sum_{i=1}^n -\log(n p_i),$$

which naturally leads to the log empirical likelihood as defined in (1.1.1)

$$\ell_{\mathrm{EL}}(\mu) := \log \mathcal{L}_{\mathrm{EL}}(\mu) = \max_{\{p_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \log p_i : \sum_{i=1}^n p_i = 1, p_i \geq 0, \sum_{i=1}^n p_i (X_i - \mu) = 0 \right\}.$$

Most importantly, [Owe90] proved the following Wilks property

$$-2\ell_{\mathrm{EL}}(\mu_0) - 2n \log n \xrightarrow{d} \chi_1^2.$$

Based on this asymptotic result, we can not only perform hypothesis testing but also construct confidence interval for the mean parameter with data driven shape.

An overview of the EL methods can be found in [Owe01] and [CVK09].

1.2 Big Data Analysis

In the age of information and technology, along with the advancement of technological revolution, information acquisition is becoming easy and cheap, which leads to the explosion of data collection through automated data collection processes. From various fields such as biomedical sciences, engineering and social sciences, massive data characterized by high dimensionality are popping up all the time. For example, with the rapid next generation sequencing technology development, hundreds of thousands of genetic variants such as single nucleotide polymorphisms (SNPs), are potential features in genome wide association studies (GWAS). Time series with very dense time points can be collected from hundreds of thousands of regions in economics, earth sciences, as well as neuroscience. In the Big Data era, documents, images, videos and other objects can all be regarded as forms of massive data. Statisticians have been also proposing new statistical methodologies to discover knowledge from those big data. For example, from studying data points in the finite Euclidean spaces to studying curves (i.e. functional data analysis), surfaces, even manifolds directly in infinite dimensional spaces.

1.2.1 Functional Data Analysis

We consider the following general functional linear regression model,

$$Y_i(t_{ij}) = \boldsymbol{\beta}_0^{\intercal}(t_{ij})\mathbf{X}_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, \cdots, n; j = 1, \cdots, m_i \tag{1.2.3}$$

where $\mathbf{X}_i(t) \sim \{\boldsymbol{\mu}(t), \boldsymbol{\Gamma}(s,t)\}$, $t_{ij} \sim f(t)$ and $\epsilon_i(t) \sim \{0, \Omega(s,t)\}$ are mutually independent. For convenience, assume that with m_i's $(1 \leq i \leq n)$ are all of the same order as $m = n^\eta$ for some $\eta \geq 0$. Data with $\eta = 0$, are called sparse functional data, i.e. longitudinal data; those satisfying $\eta \geq \eta_0$, where η_0 is a transition point to be specified, are referred as dense functional data. The scenarios with $\eta \in (0, \eta_0)$ are in a grey zone in the literature and we refer them as "moderately dense".

Historically, sparse and dense functional data were analyzed with different methodologies. For dense functional data, one can smooth each curve separately and proceed with further estimation and inference based on the pre-smoothed curves. A partial list of recent literature on dense functional data includes [CLS86], [RS91], [ZC07], [EH08] and [BHK$^+$09]. For sparse functional data, the pre-smoothing approach is not applicable and, instead, one needs to pool all data together to borrow strength from individual curves [YMW05a, YMW05b]. [HMW06] investigated the theoretical properties of functional principal component analysis based on local linear smoothers. They found that, for dense functional data with $\eta \geq 1/4$, the pre-smoothing errors are asymptotically negligible and quantities such as the mean, covariance and eigenfunctions can be estimated with a parametric root-n rate, while these quantities can only be estimated with a nonparametric convergence rate for sparse functional data with $\eta = 0$. Since sparse and dense functional data are asymptotic concepts and are hard to distinguish in reality, [LH10] proposed an estimation procedure treating all types of functional data under a unified framework including the moderately dense cases. More recently, [KZ13] proposed a unified, self-normalizing approach to construct pointwise confidence intervals for the mean function of functional data. The aforementioned papers established $\eta_0 = 1/4$ as the transition point to parametric convergence rate.

In contrast to estimation, less is known about the inference for functional data, with a

few exceptions such as [ZC07] and [KZ13]. In Chapter 2 and 3 of the thesis, we propose pointwise and simultaneous inference procedures for the functional linear model under a unified framework for all types of functional data and investigating the phase transition from sparse to dense data. We are not only the first one to propose an unified inference procedure in the regression setup which can cover all types of functional data, but also the first one to investigate the transition phase from sparse to dense functional data, for the following very broad hypothesis testing problem

$$H_0 : H\{\boldsymbol{\beta}_0(t)\} = 0 \quad \text{vs} \quad H_{1n} : H\{\boldsymbol{\beta}_0(t)\} = b_n \mathbf{d}(t) \tag{1.2.4}$$

where $H(\cdot)$ is any specified functional with some regular condition and b_n is the detectable order of local alternatives to be specified (Table 1.1). In Chapter 2 and 3, we not only derive the asymptotic distributions under the null hypothesis and local alternatives, but also propose a wild bootstrapping approach to unify the inference procedure in practice along with a nice bandwidth selection method.

Table 1.1: **Transition phase point from sparse to dense data and optimal detectable order of local alternatives for both pointwise and simultaneous inference.** Note that we lowered the transition phase point η_0 which was 1/4 in the existing literature.

	Pointwise Inference $\eta_0 = 1/8$		Simultaneous Inference $\eta_0 = 1/16$	
	$0 \le \eta < \eta_0$	$\eta \ge \eta_0$	$0 \le \eta < \eta_0$	$\eta \ge \eta_0$
b_n	$n^{-4(1+\eta)/9}$	$n^{-1/2}$	$n^{-8(1+\eta)/17}$	$n^{-1/2}$

1.2.2 High Dimensional Data Analysis

Rapid progress has been made during the past decade in high dimensional statistics, especially in linear regression model as one of the classical models in statistical theory. The vast majority of existing literature has been pursued for estimation under sparsity and homoscedasticity based on regularization with different penalties, either convex or nonconvex. The most popular representative of convex penalties is the Lasso penalty [Tib96]. The theoretical properties of the Lasso estimator such as the oracle property, which refers to consistently recovering the sparse pattern and estimating the parameters of the coefficient vector, and selection consistency have been investigated by [MY09, BRT09, BTW$^+$07, VdG08, Zha09, NRWY12] and [MB06, ZY06, Wai09]. The nonconvex representatives include SCAD [FL01], MCP [Zha10], among others. A comprehensive overview of high dimensional estimation for homoscedastic regression models can be found in [BVDG11].

Despite its prevalence in statistical data sets, heteroscedasticity has been largely ignored in high dimensional statistics literature. [WWL12] analyzed the heteroscedasticity in high dimensional case by using quantile regression. [DCL12] proposed a methodology that allows nonconstant error variances for high dimensional estimation but with a parametric form of the variance function. And recently, [BCW14] came up with a self-tuning $\sqrt{\text{Lasso}}$ estimation method that solved this important problem in high dimensional regression analysis.

Although people have made significant progress towards understanding the estimation theory for high dimensional models, very little work has been done for constructing confidence intervals, statistical testing and assigning uncertainty for penalized estimators in high dimensional sparse models. In an early work, [KF00] showed that the limiting distribution of the Lasso estimator is not normal even in the low dimensional setting. Recently, [GVHF11] and

[CG14] considered global testing with high dimensional alternative. [MMB09] and [WR09] considered p-values based on the sample splitting technique. Stability selection [MB10] and its modification [SS13] provide another procedure to estimate error measures for false positive selections in general high dimensional settings. For the lasso estimator, [LTTT14] and [TLTT14] considered an interesting conditional inference with random hypothesis, which is philosophically different with the traditional unconditional inference.

In terms of testing the significance of one single regression coefficient, the classical $z-$test (or $t-$test) is no longer applicable because the high dimensionality. People have been proposing low-dimensional projection procedure to conduct hypothesis testing and construct confidence regions [ZZ14, B$^+$13, JM13, vdGBR13, LZL$^+$13, NL14]. The way to select the projection variables varies from method to method. Some of them use node-wise Lasso procedure to select the projection variables, and some of them use the so called Key conFounder Controlling (KFC) method motivated by screening approaches [FL08].

However, all the above inference procedures assumed homoscedasticity for the error term, in particular, the conditional variance of the error is a constant. This is essential for their inference procedure to be valid since they require the accurate estimation of the error variance. Without homoscedasticity, it is hard for them to carry out the estimation of the error variance in high dimensional settings. But this hardly holds in practice. There is rarely good cause to have strong belief in the assumption that the errors are homoscedastic and similarly there is rarely sufficient information to enable accurate specification of the variance function. The use of incorrect variance models will, in general, lead to inferences that are not asymptotically valid [Bel02]. [WD12] generalized the asymptotic results of [KF00] for the case of a fixed parameter dimension under heteroscedasitic errors. But there is little work in dealing with heteroscedasticity under growing dimension along with sample size. To

bridge this gap, in Chapter 4 of this thesis, we propose to use Empirical Likelihood (EL) to test statistical hypotheses and construct confidence regions for low dimensional components in high dimensional liner models with heteroscedastic noise.

Chapter 2

Unified pointwise empirical likelihood ratio tests for functional linear models and the phase transition from sparse to dense functional data

2.1 Introduction

We consider statistical inference problems under a general functional linear regression model, where both the response $Y(t)$ and the covariate $\mathbf{X}(t) = \{X^{(1)}(t), \ldots, X^{(p)}(t)\}^\intercal$ are defined continuously on a time interval $[a, b]$. The relationship between $Y(t)$ and $\mathbf{X}(t)$ is given by

$$Y(t) = \boldsymbol{\beta}_0^\intercal(t)\mathbf{X}(t) + \epsilon(t), \tag{2.1.1}$$

where $\boldsymbol{\beta}_0(t) = \big(\beta_{10}(t), \cdots, \beta_{p0}(t)\big)^\intercal$ is a p-dimensional vector of unknown functions and $\epsilon(t)$ is a zero mean error process, independent of $\mathbf{X}$ and with a covariance function $\Omega(s,t) = \mathrm{Cov}\{\epsilon(s), \epsilon(t)\}$. The model in (2.1.1) is also referred as the concurrent functional linear model in [SR05], which includes the varying coefficient models and functional analysis of variance (fANOVA) models [MC06, ZHM+10] as special cases. In many fANOVA applications, some

components of $\mathbf{X}(t)$ are random indicators of treatment assignments with complicated cross or nested structures, see [FZ00] for more discussions on the relationship and difference between model (2.1.1) and the varying coefficient models. Without loss of generality, we allow $\mathbf{X}(t)$ to be a multivariate random process with mean function $\boldsymbol{\mu}(t) = E\{\mathbf{X}(t)\}$ and covariance function $\boldsymbol{\Gamma}(s,t) = \mathrm{Cov}\{\mathbf{X}(s), \mathbf{X}(t)\}$.

Let $\{Y_i(t), \mathbf{X}_i(t)\}$, $i = 1, \ldots, n$, be independent realizations of $\{Y(t), \mathbf{X}(t)\}$. Instead of observing the entire trajectories, one can only observe $Y_i(t)$ and $\mathbf{X}_i(t)$ on discrete time points $\{t_{ij}, j = 1 \ldots, m_i\}$. For convenience, denote $Y_{ij} = Y_i(t_{ij})$ and $X_{ij}^{(k)} = X_i^{(k)}(t_{ij})$, and assume that m_i's $(1 \leq i \leq n)$ are all of the same order as $m = n^{\eta}$ for some $\eta \geq 0$. That is m_i/m are bounded below and above by some constants. Functional data are considered to be sparse or dense depending on the order of m [HMW06, LH10]. Data with bounded m, or $\eta = 0$, are called sparse functional data; those satisfying $\eta \geq \eta_0$, where η_0 is a transition point to be specified below, are referred as dense functional data. The scenarios with $\eta \in (0, \eta_0)$ are in a grey zone in the literature and we refer them as "moderately dense" in this chapter.

As we mentioned in Section 1.2.1, sparse and dense functional data were analyzed with different methodologies. But sparse and dense functional data are asymptotic concepts and are hard to distinguish in practice, [LH10] proposed an estimation procedure treating all types of functional data under a unified framework including the moderately dense cases and they found $\eta_0 = 1/4$ is the transition point to parametric convergence rate in the estimation. In contrast to estimation, less is known about the inference for functional data, with a few exceptions such as [ZC07] and [KZ13]. The focus of the chapter is on proposing pointwise and simultaneous inference procedures for the functional linear model in (2.1.1) under a unified framework for all types of functional data and investigating the phase transition from sparse

to dense data. We are interested in testing

$$H_0 : H\{\boldsymbol{\beta}_0(t)\} = 0 \quad \text{vs} \quad H_1 : H\{\boldsymbol{\beta}_0(t)\} \neq 0 \tag{2.1.2}$$

where $H\{\mathbf{z}\}$ is a q-dimensional function of $\mathbf{z} = (z_1, \cdots, z_p)^\intercal \in \mathbb{R}^p$ such that $C(\mathbf{z}) := \frac{\partial H(\mathbf{z})}{\partial \mathbf{z}^\intercal}$ is a $q \times p$ full rank matrix ($q \leq p$) for all $\mathbf{z}$.

The test problem in (2.1.2) is very broad, including many interesting hypotheses as special cases. For instance, if $H\{\mathbf{z}\} = \mathbf{z}$, the null hypothesis is equivalent to $H_0 : \beta_{k0}(\cdot) = 0$ for all k. If $H\{\mathbf{z}\} = (z_1 - z_2, z_2 - z_3, \cdots, z_{p-1} - z_p)^\intercal$, then (2.1.2) is essentially an ANOVA hypothesis for the coefficient functions $\beta_{k0}(\cdot)$. If $H\{\mathbf{z}\} = \boldsymbol{\Lambda}\mathbf{z} - \mathbf{c}_0$ for a $q \times p$ known constant matrix $\boldsymbol{\Lambda}$ and a known vector $\mathbf{c}_0$, then (2.1.2) becomes $H_0 : \boldsymbol{\Lambda}\boldsymbol{\beta}_0(\cdot) = \mathbf{c}_0$, which is a test for linear constraints on $\boldsymbol{\beta}_0(\cdot)$. Similar hypothesis testing problems have been studied by [ZC07] and [Zha11]. However, their methods only apply to dense functional data with $\eta > 5/4$.

In this chapter, we propose nonparametric tests based on the empirical likelihood (EL) to test (2.1.2) pointwisely. We show the EL-based tests enjoy a nice self-normalizing property such that we can treat both sparse and dense functional data under a unified framework. There have been some works on EL methods for sparse functional data with $\eta = 0$. Among them, [XZ07] proposed an EL method for constructing pointwise confidence interval and a Bonferroni type simultaneous confidence band for the mean function. [CZ10] studied an EL-based method for testing ANOVA type hypotheses in partial linear models with missing values.

To investigate the power of the tests, we consider the local alternatives

$$H_{1n} : H\{\boldsymbol{\beta}_0(t)\} = b_n \mathbf{d}(t), \tag{2.1.3}$$

where b_n is a sequence of numbers converging to 0 at a rate to be specified later and $\mathbf{d}(t) \neq 0$ is any q-dimensional function. For a given test, b_n is the smallest order of the local alternatives so that the test has a non-trivial power for any fixed non-zero $\mathbf{d}(\cdot)$. Thus b_n quantifies the order of signals that a test can detect. For the sparse data with $\eta = 0$, it is known that the EL method using a global bandwidth h [CZ10] can detect alternatives of order $b_n = (nh)^{-1/2}$ for pointwise tests. Since $h \to 0$ in a typical nonparametric regression setting, the detectable order here is larger than $n^{-1/2}$. However, for dense data with $\eta > 0$, the detectable order b_n is still largely unknown. One key interest in this chapter is to understand the effect of η on b_n in the pointwise test. The optimal b_n is obtained by maximizing the power of the test (i.e., minimizing the order of b_n) while controlling the type I error at the desired level. Under some mild conditions, we find that, for the pointwise test, the optimal rate b_n is larger than $n^{-1/2}$ for $\eta \leq 1/8$ and equals to $n^{-1/2}$ for $\eta > 1/8$. The transition point $1/8$ will be refereed as η_0 for the pointwise tests. Once $\eta > \eta_0$, with a properly chosen bandwidth, the proposed tests can detect a signal at a parametric rate.

The rest of the chapter is organized as follows. In Section 2.2, we present a bias-corrected estimator and some preliminary results. We propose the unified pointwise EL test in Section 2.3 where we investigate the asymptotic distributions of the test statistic under both the null and local alternatives, and the transition phases for b_n. In Section 2.4, we address implementation issues such as bandwidth selection and covariance estimation. Simulation studies are presented in Section 2.5. All the technical details are relegated to the Section 2.6.

2.2 A bias-corrected estimator and some preliminary results

In this section, we will first introduce an initial local linear estimator $\hat{\boldsymbol{\beta}}(t)$ [FG96] for $\boldsymbol{\beta}_0(t)$ and then introduce a bias-corrected estimator $\check{\beta}(t)$ and some preliminary results.

2.2.1 A bias-corrected estimator

Let $K(\cdot)$ be a symmetric probability density function that we use as a kernel, h be a bandwidth, and denote $K_h(\cdot) = K(\cdot/h)/h$. For any t in a neighborhood of t_0, $\beta_{k0}(t)$ can be approximated by

$$\beta_{k0}(t) \approx \beta_{k0}(t_0) + \frac{\partial \beta_{k0}(t_0)}{\partial t}(t - t_0) := a_k + b_k(t - t_0), k = 1, 2, \cdots, p.$$

Denote $\boldsymbol{\vartheta} = (a_1, \ldots, a_p, hb_1, \ldots, hb_p)^\intercal$ and $\mathbf{D}_{ij}(t) = (\mathbf{X}_{ij}^\intercal, \frac{t_{ij}-t}{h}\mathbf{X}_{ij}^\intercal)^\intercal$. Put

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^\intercal, \mathbf{Y} = (\mathbf{Y}_1^\intercal, \mathbf{Y}_2^\intercal, \cdots, \mathbf{Y}_n^\intercal)^\intercal,$$

$$\mathbf{D}_i(t) = (\mathbf{D}_{i1}(t), \mathbf{D}_{i2}(t), \cdots, \mathbf{D}_{im_i}(t))^\intercal, \mathbf{D}(t) = (\mathbf{D}_1^\intercal(t), \mathbf{D}_2^\intercal(t), \cdots, \mathbf{D}_n^\intercal(t))^\intercal,$$

$$\mathbf{W}_i(t) = \frac{1}{m_i}\mathrm{diag}\{K_h(t_{i1} - t), K_h(t_{i2} - t), \ldots, K_h(t_{im_i} - t)\},$$

$$\text{and } \mathbf{W}(t) = \mathrm{diag}(\mathbf{W}_1(t), \mathbf{W}_2(t), \ldots, \mathbf{W}_n(t)).$$

An estimator for $\boldsymbol{\vartheta}$ is obtained as

$$\begin{aligned} \hat{\boldsymbol{\vartheta}} &= \arg\min_{\boldsymbol{\vartheta}} [\mathbf{Y} - \mathbf{D}(t_0)\boldsymbol{\vartheta}]^\intercal \mathbf{W}(t_0)[\mathbf{Y} - \mathbf{D}(t_0)\boldsymbol{\vartheta}], \\ &= [\mathbf{D}^\intercal(t_0)\mathbf{W}(t_0)\mathbf{D}(t_0)]^{-1}\mathbf{D}^\intercal(t_0)\mathbf{W}(t_0)\mathbf{Y}. \end{aligned} \tag{2.2.4}$$

Thus the local linear estimator for $\boldsymbol{\beta}_0(t_0)$ is

$$\hat{\boldsymbol{\beta}}(t_0) = \big(\mathbf{I}_p, \mathbf{0}_p\big)\hat{\boldsymbol{\vartheta}} = \big(\mathbf{I}_p, \mathbf{0}_p\big)[\mathbf{D}^\intercal(t_0)\mathbf{W}(t_0)\mathbf{D}(t_0)]^{-1}\mathbf{D}^\intercal(t_0)\mathbf{W}(t_0)\mathbf{Y}, \tag{2.2.5}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\mathbf{0}_p$ is a $p \times p$ zero matrix. It is shown in Lemma 1 in Section 2.6.2 that

$$\sup_{t\in[a,b]} |\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)| = O\left\{h^2 + (\frac{\log n}{n} + \frac{\log n}{nmh})^{1/2}\right\} \quad a.s. \tag{2.2.6}$$

Since the bias of $\hat{\boldsymbol{\beta}}(t)$ is of order h^2, undersmoothing is typically needed for an unbiased testing procedure based on $\hat{\boldsymbol{\beta}}(t)$ [XZ07]. To avoid undersmoothing and reduce the estimation bias in $\hat{\boldsymbol{\beta}}(t)$, we define $\check{\boldsymbol{\beta}}(t)$ as the solution of the following residual-adjusted [XZ07] estimating equation for $\boldsymbol{\beta}(t)$

$$\bar{g}_n\{\boldsymbol{\beta}(t)\} := \frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\} = 0, \tag{2.2.7}$$

with $g_i\{\boldsymbol{\beta}(t)\} = \frac{1}{m_i}\sum_{j=1}^{m_i}\left\{Y_{ij} - \boldsymbol{\beta}^\intercal(t)\mathbf{X}_{ij} - \{\hat{\boldsymbol{\beta}}(t_{ij}) - \hat{\boldsymbol{\beta}}(t)\}^\intercal\mathbf{X}_{ij}\right\}\mathbf{X}_{ij}K_h(t_{ij}-t)$, where $\hat{\boldsymbol{\beta}}(t)$ is the local linear estimator for $\boldsymbol{\beta}_0(t)$.

2.2.2 Regularity conditions and preliminary results

We now present some preliminary results regarding the asymptotics of $\check{\boldsymbol{\beta}}(t)$. Assume that t_{ij} are i.i.d. random variables following a probability density function $f(t)$. For convenience, define $\boldsymbol{\Gamma}(t) = \boldsymbol{\Gamma}(t,t)$, $\Omega(t) = \Omega(t,t)$, $\mathbf{C}(t) = C\{\boldsymbol{\beta}_0(t)\}$ and $\mathbf{A}(t) = \boldsymbol{\Gamma}(t)f(t)$. We will also use $\tilde{o}_p$ and $\tilde{O}_p$ to represent that, respectively, o_p and O_p hold uniformly for all $t \in [a,b]$. The

following conditions are needed for our asymptotic results.

(C1): The kernel function $K(\cdot)$ is a symmetric probability density function with a bounded support in $[-1, 1]$.

(C2): Assume that $\mathbb{E}\left\{\sup_{t\in[a,b]} \|\mathbf{X}(t)\|^{\lambda_1}\right\} < \infty$ and $\mathbb{E}\left\{\sup_{t\in[a,b]} |\epsilon(t)|^{\lambda_2}\right\} < \infty$ for some $\lambda_1, \lambda_2 \geq 5$ where $\|\cdot\|$ is the L_2 norm for a vector.

(C3): Assume that $f(t)$ and $\mathbf{\Gamma}(t)$ have continuous second derivatives on $[a, b]$, $\boldsymbol{\beta}_0(t)$ has continuous third derivatives on $t \in [a, b]$, and $\mathbf{C}(t)$ is uniformly bounded on $t \in [a, b]$.

(C4): Define $\lambda = \min\{\lambda_1, \lambda_2\}$ and let $h = n^{-\alpha_0}$ with $\alpha_0 \in (0, 1)$ being the order of the bandwidth. Assume that (i) $\alpha_0 < 1 - \eta - 2/\lambda$ if $\eta \in [0, 1/8]$ and $\alpha_0 < 1/2 - 1/\lambda$ if $\eta > 1/8$; (ii) $(1 + \eta)/9 < \alpha_0$ if $\eta \in [0, 1/8]$ and $1/8 < \alpha_0 < \eta$ if $\eta > 1/8$.

Conditions (C1) and (C3) are commonly used regularity conditions in nonparametric regressions. Condition (C2) is similar to that in [LH10]. The upper bounds on the bandwidth h in (C4)(i) are adapted from [LH10]. Detailed explanation on the restrictions on h in (C4)(ii) will be given in Remark 2 after Proposition 2. Selecting a bandwidth that satisfies (C4) will be discussed in Section 2.4.

The following proposition provides an asymptotic expansion for $\check{\boldsymbol{\beta}}(t)$.

Proposition 1. *Under conditions (C1)-(C3) and (C4)(i),*

$$\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) = -\mathbf{A}^{-1}(t)\bar{\boldsymbol{\xi}}_n(t)\{1 + \tilde{o}_p(1)\} + \tilde{O}_p(h^4), \tag{2.2.8}$$

where $\bar{\boldsymbol{\xi}}_n(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)$ *and* $\xi_i(t) = m_i^{-1}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)$. *Let*

$$\bar{r} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} m/m_i, \mu_{ts} = \int u^s K^t(u)du,$$

then

$$Var\{\bar{\boldsymbol{\xi}}_n(t)\} = \boldsymbol{\Gamma}(t)\Omega(t)f(t)\Big\{\frac{\bar{r}}{mnh}\mu_{20} + \frac{m-\bar{r}}{nm}f(t)\Big\}\{1+\tilde{o}(1)\}. \tag{2.2.9}$$

The proof of Proposition 1 is provided in Section 2.6.2.

Remark 1. *Proposition 1 implies that the mean square error (MSE) of* $\check{\boldsymbol{\beta}}(t)$ *is* $MSE\left\{\check{\boldsymbol{\beta}}(t)\right\} = O\{h^8 + \frac{1}{mnh} + \frac{1}{n}\}$. *Hence the optimal* h_{opt} *that minimize the MSE of* $\check{\boldsymbol{\beta}}(t)$ *is* $h_{opt} \sim (mn)^{-1/9} = n^{-(1+\eta)/9}$. *It follows that*

$$\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) = O_p\{h_{opt}^4 + (mnh_{opt})^{-\frac{1}{2}} + n^{-\frac{1}{2}}\} = O_p\{n^{-1/2} + n^{-4(1+\eta)/9}\}.$$

Then the optimal convergence rate of $\check{\boldsymbol{\beta}}(t)$ *is of order* $n^{-4(1+\eta)/9}$ *if* $\eta \le 1/8$ *and of order* $n^{-1/2}$ *if* $\eta > 1/8$. *Thus,* $\eta_0 = 1/8$ *is the transition point for the convergence rate of* $\check{\boldsymbol{\beta}}(t)$. *When* $\eta > \eta_0$, $\check{\boldsymbol{\beta}}(t)$ *is no longer sensitive to the choice of* h *and its the convergence rate remains at* $O_p(n^{-1/2})$ *as long as* $h = O(n^{-1/8})$ *and* $h \gg m^{-1} = n^{-\eta}$.

The following proposition provides the asymptotic distribution of $\check{\boldsymbol{\beta}}(t)$ and its proof is provided in Section 2.6.2.

Proposition 2. *Suppose* $mh \to \kappa_0 \in [0,\infty]$, *define*

$$C_{n,\alpha_0,\eta} = \begin{cases} \{n/(mh)\}^{1/2}, & \text{if } \kappa_0 < \infty; \\ n^{1/2}, & \text{if } \kappa_0 = \infty \end{cases} \tag{2.2.10}$$

and $\mathbf{B}(t) = \mathbf{\Gamma}(t)\Omega(t)f(t)\left\{(\bar{r}\mu_{20} + \kappa_0 f(t))I(\kappa_0 < \infty) + f(t)I(\kappa_0 = \infty)\right\}$. *Under conditions (C1)-(C4), we have*

$$nC_{n,\alpha_0,\eta}^{-1}\left\{\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\right\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}(t)). \tag{2.2.11}$$

where $\mathbf{V}(t) = \mathbf{A}^{-1}(t)\mathbf{B}(t)\mathbf{A}^{-1}(t)$.

Remark 2. *By Proposition 1, the bias in* $nC_{n,\alpha_0,\eta}^{-1}\{\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\}$ *is of order* $O_p(nh^4/C_{n,\alpha_0,\eta})$. *Since the bias can lead to invalid tests, we use Condition (C4) (ii) to ensure that the bias is asymptotically negligible. When* $\eta \leq \eta_0 = 1/8$, *the condition* $\alpha_0 > (1+\eta)/9$ *warrants that* $mh < \infty$ *and hence* $nh^4/C_{n,\alpha_0,\eta} = n^{1/2}m^{1/2}h^{9/2} = n^{(1+\eta-9\alpha_0)/2} = o(1)$. *When* $\eta > \eta_0$, *the condition that* $1/8 < \alpha_0 < \eta$ *implies* $mh \to \infty$ *and* $nh^4/C_{n,\alpha_0,\eta} = n^{1/2}h^4 = n^{1/2-4\alpha_0} \to 0$.

By Proposition 2 and the Delta method, we can show that, under H_0,

$$nC_{n,\alpha_0,\eta}^{-1}H\{\check{\boldsymbol{\beta}}(t)\} \xrightarrow{d} N(\mathbf{0}, \mathbf{R}^{-1}(t)) \tag{2.2.12}$$

where $\mathbf{R}(t) = \{\mathbf{C}(t)\mathbf{V}(t)\mathbf{C}(t)^\intercal\}^{-1}$. The asymptotic variances of $H\{\check{\boldsymbol{\beta}}(t)\}$ are different under sparse and dense cases. A Wald-type test statistic may be constructed using (2.2.12) if an appropriate estimator for the variance of $H\{\check{\boldsymbol{\beta}}(t)\}$ can be obtained. But we will not pursue this direction because the estimation of the asymptotic variance involves many nonparametric functions e.g. $\mathbf{\Gamma}(t), \Omega(t)$ and $f(t)$, which requires properly selecting several bandwidths. Instead, we propose a self-normalizing EL method in the next section which avoids estimating the asymptotic variance explicitly.

2.3 A unified pointwise test

In this section, we will introduce a unified test for H_0 at any fixed time t, which is based on the empirical likelihood ratio (ELR) statistic. To construct an ELR statistic for testing (2.1.2), we first define the EL function at $\boldsymbol{\beta}(t)$ for a fixed $t \in [a, b]$. Following [Owe90], the empirical likelihood for $\boldsymbol{\beta}(t)$ is defined as

$$L\{\boldsymbol{\beta}(t)\} = \max_{p_1, p_2, \ldots, p_n} \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, p_i \geq 0, \sum_{i=1}^{n} p_i g_i\{\boldsymbol{\beta}(t)\} = 0 \right\}.$$

Applying the Lagrange multiplier, the log-EL function becomes

$$l\{\boldsymbol{\beta}(t)\} := \log L\{\boldsymbol{\beta}(t)\} = -\sum \log\{1 + \boldsymbol{\gamma}^{\intercal}(t) g_i\{\boldsymbol{\beta}(t)\}\} - n \log n$$

where $\boldsymbol{\gamma}(t)$ is a solution to the following equation

$$Q_{1n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t)\} := \frac{1}{n} \sum_{i=1}^{n} \frac{g_i\{\boldsymbol{\beta}(t)\}}{1 + \boldsymbol{\gamma}^{\intercal}(t) g_i\{\boldsymbol{\beta}(t)\}} = 0. \tag{2.3.13}$$

The maximum log-EL without any constraint is $l\{\check{\boldsymbol{\beta}}(t)\} = -n \log n$. It follows that the negative log-ELR for testing $H_0 : H\{\boldsymbol{\beta}_0(t)\} = 0$ is

$$\ell(t) := \min_{H\{\boldsymbol{\beta}(t)\}=0} l_0\{\boldsymbol{\beta}(t)\}, \tag{2.3.14}$$

where $l_0\{\boldsymbol{\beta}(t)\} = \sum_{i=1}^{n} \log\{1 + \boldsymbol{\gamma}^{\intercal}(t) g_i\{\boldsymbol{\beta}(t)\}\}$. To solve (2.3.14), we minimize the following objective function [QL95]

$$M\{\boldsymbol{\beta}(t), \boldsymbol{\nu}(t)\} = \frac{1}{n} l_0\{\boldsymbol{\beta}(t)\} + \boldsymbol{\nu}^{\intercal}(t) H\{\boldsymbol{\beta}(t)\},$$

where $\boldsymbol{\nu}(t)$ is a $q \times 1$ vector of Lagrange multipliers. Differentiating $M(\cdot, \cdot)$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ and setting them to zero, we have

$$Q_{2n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)\} := \frac{1}{n}\frac{\partial l_0\{\boldsymbol{\beta}(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} + C^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t) = 0 \text{ and } H\{\boldsymbol{\beta}(t)\} = 0.$$

Combining equation (2.3.13) for $\boldsymbol{\gamma}(t)$, the constrained minimization problem in (2.3.14) is equivalent to solving the following estimating equation system

$$Q_{1n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t)\} = 0; \quad Q_{2n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)\} = 0 \text{ and } H\{\boldsymbol{\beta}(t)\} = 0. \tag{2.3.15}$$

We show in Section 2.6.2.3 that a consistent solution to (2.3.15), denoted as $(\tilde{\boldsymbol{\beta}}(t), \tilde{\boldsymbol{\gamma}}(t), \tilde{\boldsymbol{\nu}}(t))$, exists almost surely. We call $\tilde{\boldsymbol{\beta}}(t)$ the Restricted Maximum Empirical Likelihood Estimator (RMELE). Then the test statistic in (2.3.14) becomes

$$\ell(t) = l_0\{\tilde{\boldsymbol{\beta}}(t)\}. \tag{2.3.16}$$

The following proposition provides an asymptotic expansion for $2\ell(t)$.

Proposition 3. *Under conditions (C1)-(C4), and under H_0, we have, for each $t \in [a, b]$,*

$$2\ell(t) = \mathbf{U}_n(t)^{\mathsf{T}}\mathbf{U}_n(t) + O_p(nh^4/C_{n,\alpha_0,\eta}), \tag{2.3.17}$$

where $\mathbf{U}_n(t) = nC_{n,\alpha_0,\eta}^{-1}\mathbf{G}(t)\bar{\boldsymbol{\xi}}_n(t)$, $\mathbf{G}(t) = \mathbf{R}^{1/2}(t)\mathbf{C}(t)\mathbf{A}^{-1}(t)$ and $\mathbf{R}(t)$ and $\mathbf{A}(t)$ are the same as defined in (2.2.12).

The asymptotic expansion in (2.3.17) makes a connection between $2\ell(t)$ and the bias-corrected estimator $\check{\boldsymbol{\beta}}(t)$ described in Section 2.2. By Proposition 1 and (2.2.12), $\mathbf{U}_n(t) =$

$nC_{n,\alpha_0,\eta}^{-1}\mathbf{R}^{1/2}(t)H\{\check{\boldsymbol{\beta}}(t)\} + o_p(1)$ and asymptotically follows a q-dimensional multivariate standard normal distribution. Naturally, $2\ell(t) \xrightarrow{d} \chi_q^2$ under the null hypothesis. The fact that the asymptotic distribution of $2\ell(t)$ does not depend on m (or η) proves that it is a self-normalized test statistic no matter the data are sparse or dense. This is a very appealing property because the test procedure is the same for all types of functional data and solving (2.3.15) does not require estimating the variance of $H\{\check{\boldsymbol{\beta}}(t)\}$.

The following Theorem summarizes the asymptotic distribution of $2\ell(t)$ under both H_0 and the local alternative (2.1.3).

Theorem 1. *Under conditions (C1)-(C4) and suppose* $H\{\boldsymbol{\beta}_0(t)\} = b_n\mathbf{d}(t)$ *for* $t \in [a, b]$, *where* $b_n = n^{-1}C_{n,\alpha_0,\eta}$ *and* $\mathbf{d}(t)$ *is any fixed real vector of functions, we have*

$$2\ell(t) \xrightarrow{d} \chi_q^2\{\mathbf{d}^\intercal(t)\mathbf{R}(t)\mathbf{d}(t)\}$$

where $\mathbf{d}^\intercal(t)\mathbf{R}(t)\mathbf{d}(t)$ *is the noncentrality parameter.*

A proof of Theorem 1 is provided in the Section 3.5.1.

Remark 3. *Under* H_0, $\mathbf{d}(t) = 0$ *and Theorem 1 suggests that* $2\ell(t)$ *follows a* χ_q^2 *distribution asymptotically. An asymptotic* α *level test is given by rejecting* H_0 *at a fixed point* t *if* $2\ell(t) > \chi_{q,\alpha}^2$ *where* $\chi_{q,\alpha}^2$ *is the upper* α *quantile of* χ_q^2. *By taking a special function* $H\{\boldsymbol{\beta}\} = \beta_j(t)$, *we can also construct a* $(1-\alpha)100\%$ *confidence interval for* $\beta_j(t)$ $(j = 1, \cdots, p)$ *as* $CI_\alpha = \{\beta_j(t) : 2\ell(t) < \chi_{q,\alpha}^2\}$, *which can be computed numerically. This provides an alternative self-normalized confidence interval to those based on a self-normalized normal approximation [KZ13]. Comparing to Kim and Zhao's method, our method does not require estimating the bias because we use bias-corrected estimating equations.*

We define the size of the detectable signal b_n^* as the smallest order b_n in (2.1.3) that the proposed test can detect. For a given significant level α,

$$b_n^* = \min_h b_n \text{ subject to (i) Type I error} \leq \alpha \text{ under } H_0 \qquad (2.3.18)$$
$$\text{and (ii) the power is non-trivial under } H_{1n}.$$

Theorem 1 guarantees that the proposed test controls the Type I error at the nominal level asymptotically. For the sparse and moderate dense cases ($\eta \leq 1/8$), Condition (C4) implies $mh \to 0$ and hence $b_n = (nmh)^{-1/2}$ by Theorem 1. In this case, b_n^* is equivalent to

$$\min_h b_n = (nmh)^{-1/2} \text{ subject to condition (C4) on } h.$$

The optimal h that solves the minimization problem above is $h_* = n^{-(1+\eta+\delta)/9}$ for an arbitrarily small $\delta > 0$. This implies the optimal b_n is $n^{-4(1+\eta)/9+\delta/18}$, which results in $b_n^* = n^{-4(1+\eta)/9}$ by letting $\delta \to 0$. For dense data ($\eta > 1/8$), (C4) leads to $mh \to \infty$. Theorem 1 implies that the proposed test has a non-trivial power under a local alternative of size $b_n^* = n^{-1/2}$, which is the detectable order of a parametric test.

2.4 Implementation issues

2.4.1 Bandwidth selection

The performance of the estimation and test procedures depends on the bandwidth h and our asymptotic theory relies on h falling in the range defined in Condition (C4). For longitudinal data (sparse functional data) where subjects are assumed to be independent, one may apply

a "leave-one-out" cross-validation strategy [RS91] to choose bandwidth. However, cross-validation is time-consuming and in general, its performance for dense functional data is unknown.

We propose to select the bandwidth through minimizing the conditional integrated mean squared error (IMSE) of the local polynomial estimator $\hat{\boldsymbol{\beta}}(t)$. By (2.2.5), the bandwidth h that minimizing the IMSE of $\hat{\boldsymbol{\beta}}(t)$ is at the order of $n^{-(1+\eta)/5}$, which satisfies condition (C4) for both sparse and dense cases. Let $\mathcal{D} = \{(t_{ij}, \mathbf{X}_{ij}),\ j = 1, 2, \cdots, m_i,\ i = 1, 2, \cdots, n\}$. It is not difficult to show that for any fixed t,

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D}) = \mathbf{b}^{\intercal}(t)\mathbf{b}(t) + \mathrm{tr}\{\mathrm{Cov}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})\}$$

where $\mathbf{b}(t) = \mathrm{Bias}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})$. The IMSE is defined as

$$\mathrm{IMSE}(\hat{\boldsymbol{\beta}}(\cdot)|\mathcal{D}) = \int_a^b \mathrm{MSE}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})\varpi(t)f(t)dt$$

where $\varpi(t)$ is a known weight function and $f(t)$ is the probability density function of t_{ij}. The conditional bias is

$$\mathbf{b}(t) = (\mathbf{I}, \mathbf{0})(\mathbf{D}^{\intercal}(t)\mathbf{W}(t)\mathbf{D}(t))^{-1}\mathbf{D}^{\intercal}(t)\mathbf{W}(t)\mathbf{l}(t),$$

where $\mathbf{l}(t) = (l_{11}(t), \cdots, l_{1m_1}(t), l_{21}(t), \cdots, l_{nm_n}(t))^{\intercal}$ with

$$\begin{aligned} l_{ij}(t) &= \mathbf{X}_{ij}^{\intercal}\boldsymbol{\beta}(t_{ij}) - \mathbf{X}_{ij}^{\intercal}[\boldsymbol{\beta}_0(t) + (t_{ij} - t)\boldsymbol{\beta}^{(1)}(t)] \\ &= \mathbf{X}_{ij}^{\intercal}[\boldsymbol{\beta}(t_{ij}) - \boldsymbol{\beta}_0(t) - (t_{ij} - t)\boldsymbol{\beta}^{(1)}(t)] \approx \mathbf{X}_{ij}^{\intercal}\boldsymbol{\beta}^{(2)}(t)(t_{ij} - t)^2/2, \end{aligned}$$

and $\boldsymbol{\beta}^{(s)}(t) = \{\beta_1^{(s)}(t), \cdots, \beta_p^{(s)}(t)\}^\intercal$, $s = 1, 2$, is the s-th derivative of $\boldsymbol{\beta}_0(t)$. The conditional covariance is

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D}) = (\mathbf{I}, \mathbf{0})(\mathbf{D}^\intercal(t)\mathbf{W}(t)\mathbf{D}(t))^{-1}\mathbf{D}^\intercal(t)\mathbf{W}(t)\boldsymbol{\Omega} \\ \times \mathbf{W}(t)\mathbf{D}(t)(\mathbf{D}^\intercal(t)\mathbf{W}(t)\mathbf{D}(t))^{-1}\begin{pmatrix}\mathbf{I} \\ \mathbf{0}\end{pmatrix},\end{aligned}$$

where $\boldsymbol{\Omega} = \mathrm{Cov}(\mathbf{Y}|\mathcal{D}) = \mathrm{diag}(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \cdots, \boldsymbol{\Omega}_n)$ and $\boldsymbol{\Omega}_i = \Big(\Omega(t_{ij}, t_{ik})\Big)_{j,k=1}^{m_i}$.

An estimator of the covariance $\Omega(s,t)$ is described in Section 2.4.2. To estimate $\boldsymbol{\beta}^{(2)}(t)$, we use a higher order local polynomial estimator of $\boldsymbol{\beta}_0(t)$ with a pilot bandwidth h^*. The pilot bandwidth is obtained by minimizing the residual squares criterion in [ZL00]. By replacing $\boldsymbol{\beta}^{(2)}(t)$ and $\boldsymbol{\Omega}$ with their estimators $\widehat{\boldsymbol{\beta}^{(2)}}(t)$ and $\hat{\boldsymbol{\Omega}}$, we obtain estimators of the conditional mean and covariance, $\hat{\mathbf{b}}(t)$ and $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})$. Then the bandwidth h is chosen by minimizing the empirical IMSE

$$\hat{h} = \arg\min_h \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\widehat{\mathrm{MSE}}\{\hat{\boldsymbol{\beta}}(t_{ij})|\mathcal{D}\}\varpi(t_{ij})$$

where $N = \sum_{i=1}^n m_i$ and $\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D}) = \hat{\mathbf{b}}^\intercal(t)\hat{\mathbf{b}}(t) + \mathrm{tr}\{\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}(t)|\mathcal{D})\}$.

2.4.2 Covariance Estimation

The covariance function $\Omega(\cdot,\cdot)$ can be estimated by the nonparametric kernel estimator of [YMW05a], which is uniformly consistent [LH10]. However, the nonparametric covariance estimator is not necessarily positive semi-definite. Instead, we adopt the semiparametric covariance estimation of [FHL07]. Suppose the covariance function can be decomposed as

$\Omega(s,t) = \sigma(s)\rho(s,t)\sigma(t)$, we model the variance function $\sigma^2(t)$ nonparametrically and the correlation function $\rho(s,t)$ parametrically. For estimation, we first apply the nonparametric kernel estimators of $\Omega(s,t)$ and $\sigma^2(t)$ [YMW05a] to get information about the parametric structure of $\rho(s,t)$. Then we fit a parametric model to $\rho(s,t)$ using the quasi maximum likelihood estimator of [FHL07]. The parametric structure guarantees the positive semi-definiteness of the estimated correlation function. For more details of the implementation, see Section 2.5.

2.5 Simulation studies

Simulation studies were conducted to evaluate the performance of the proposed unified inference procedures. We generated data from the following model

$$Y_i(t_{ij}) = \beta_1(t_{ij})X_i^{(1)}(t_{ij}) + \beta_2(t_{ij})X_i^{(2)}(t_{ij}) + \epsilon_i(t_{ij}) \tag{2.5.19}$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$ where t_{ij}'s are IID Unif[0,1] distributed, $X_i^{(1)}(t_{ij}) = 1+2e^{t_{ij}}+v_{ij}$ and $X_i^{(2)}(t_{ij}) = 3-4t_{ij}^2+u_{ij}$. Here u_{ij} and v_{ij} are IID $N(0,1)$ random variables, which are independent with t_{ij} and $\epsilon_i(t_{ij})$. The random error $\epsilon_i(t_{ij})$ was generated from a zero mean AR(1) process such that $\text{Var}\{\epsilon(t)\} = 1$ and $\text{Cov}\{\epsilon(t), \epsilon(t-s)\} = \rho^{10s}$ for some $\rho \in (0,1)$. To evaluate the proposed methods for both sparse and dense data, we set $m = 5, 10$ and 50. The sample sizes were chosen to be 100 and 200. The Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)_+$ was used for estimation, where $(a)_+ = \max(a, 0)$. Bandwidth selection was conducted for every simulated data set using the method proposed in Section 2.4.

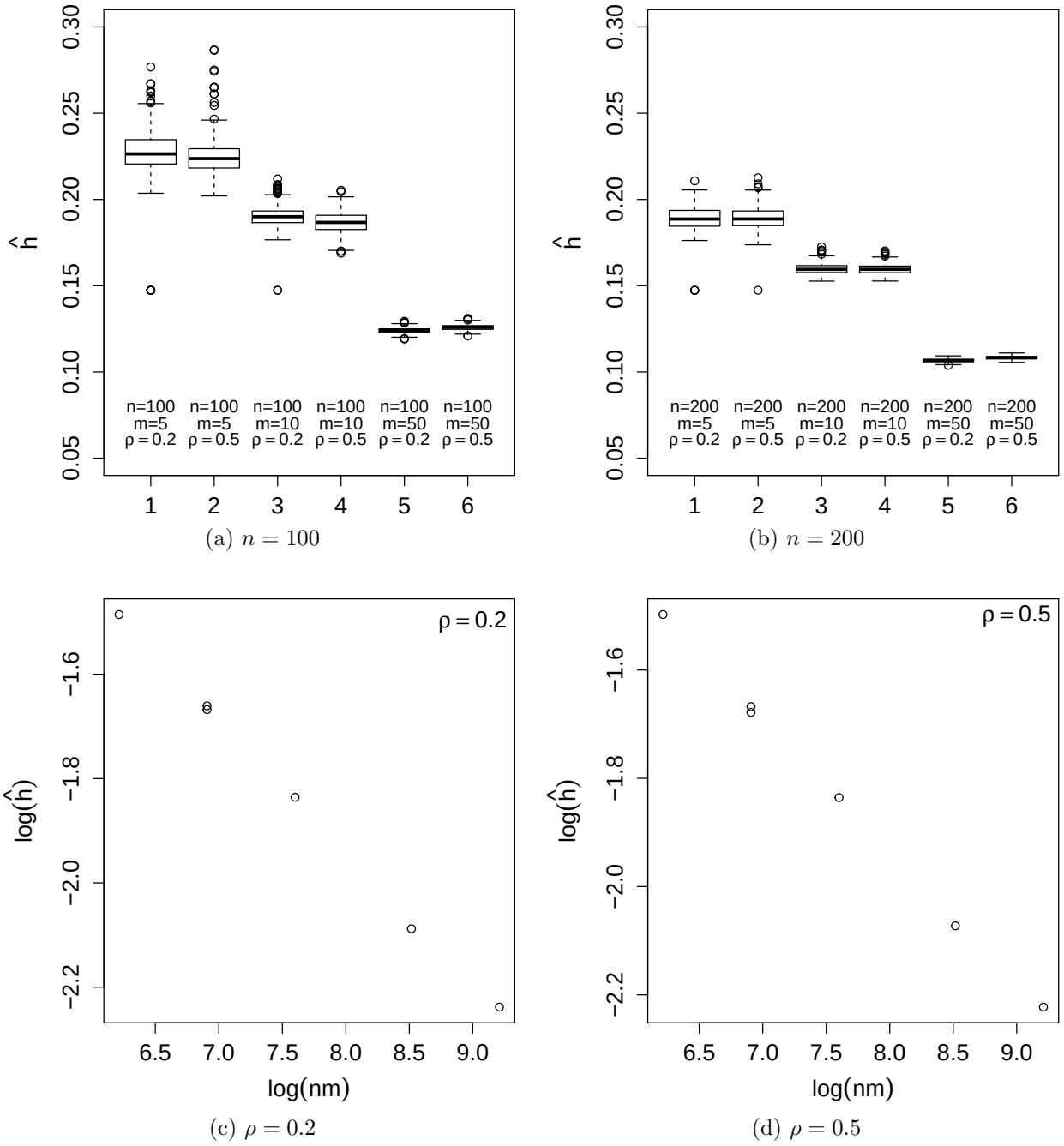
We first set $\beta_1(t) = \frac{1}{2}\sin t$ and $\beta_2(t) = 2\sin(t + 0.5)$ in Model (2.5.19) and applied the procedure in Section 2.3 to construct pointwise CIs for $\beta_1(t)$. Table 2.1 summarizes the empirical coverage probability (CP) in percentage and the average length (AL) of the CIs (in parentheses) for $\beta_1(t)$ at $t = 0.3, 0.5$ and 0.7 based on 1000 simulation replicates. These results were obtained using the data-driven bandwidth. As we can see from the table, the CPs are close to the nominal level 95% in both sparse and dense cases and the ALs are shorter under a larger sample size. In addition, the ALs improve as m increases from 5 to 50.

Table 2.1: Empirical coverage probability (%) and average length of pointwise confidence intervals (in parenthesis) for $\beta_1(t)$ at $t = 0.3, 0.5$ and 0.7.

		$m = 5$		$m = 10$		$m = 50$	
t	n	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.2$	$\rho = 0.5$
0.3	100	92.1(0.272)	92.9(0.268)	92.9(0.203)	92.5(0.203)	93.7(0.107)	93.9(0.107)
	200	92.3(0.205)	92.3(0.205)	93.5(0.152)	93.0(0.152)	94.7(0.081)	94.4(0.081)
0.5	100	92.9(0.270)	93.5(0.267)	94.5(0.210)	94.0(0.209)	93.3(0.107)	93.1(0.108)
	200	93.6(0.201)	93.3(0.200)	94.6(0.152)	94.4(0.152)	94.0(0.083)	93.8(0.081)
0.7	100	92.1(0.273)	92.5(0.272)	92.2(0.211)	92.1(0.208)	93.4(0.106)	92.8(0.106)
	200	92.3(0.201)	92.4(0.201)	94.1(0.153)	93.3(0.153)	93.9(0.083)	93.8(0.081)

To further demonstrate the performance of the proposed bandwidth selection method in Section 2.4, we show in panels (a) and (b) of Figure 2.1 the box plots of $\hat{h}$ selected for model (2.5.19) with $\beta_1(t) = \frac{1}{2}\sin(\pi t)$ and $\beta_2(t) = 2\sin(\pi t + 0.5)$ based on 500 replicates. Both the median and spread of $\hat{h}$ decreased as the n and m increased and the correlation ρ had little impact on the bandwidth selection result. These plots also show that our bandwidth selection procedure is very stable as there are very few outliers in each case. In panels (c) and (d) of Figure 2.1, we plot the logarithm of Median($\hat{h}$) against the logarithm of nm for each value of ρ. These plots show clear linear decreasing trends, confirming the selected

bandwidth decreases in a polynomial order of nm.



(a) $n = 100$ (b) $n = 200$

(c) $\rho = 0.2$ (d) $\rho = 0.5$

Figure 2.1: Panels (a) and (b) are box plots for bandwidths selected for model (2.5.19) with $\beta_1(t) = \frac{1}{2}\sin(\pi t)$ and $\beta_2(t) = 2\sin(\pi t + 0.5)$ using the proposed bandwidth selection method in Section 2.4. Panels (c) and (d) are the plots of the logarithm of median($\hat{h}$) vs log(nm).

2.6 Technical Details

This section contains the proofs for the main theorems in Section 2.3. Proofs for the propositions can be found in the next section.

2.6.1 Proof of Theorem 1

Proof of Theorem 1. For convenience, we suppress the argument of all the functions on t. Define

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} -\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{A}\mathbf{P}\mathbf{A}\mathbf{B}^{-1} & \mathbf{B}^{-1}\mathbf{A}\mathbf{P} & \mathbf{B}^{-1}\mathbf{A}\mathbf{Q}^{\intercal} \\ \mathbf{P}\mathbf{A}\mathbf{B}^{-1} & \mathbf{P} & \mathbf{Q}^{\intercal} \\ \mathbf{Q}\mathbf{A}\mathbf{B}^{-1} & \mathbf{Q} & -\mathbf{R} \end{pmatrix},$$

where $\mathbf{P} = \mathbf{V}(\mathbf{I} - \mathbf{C}^{\intercal}\mathbf{Q})$ and $\mathbf{Q} = \mathbf{R}\mathbf{C}\mathbf{V}$. By Taylor expansion of the equations (2.3.15) at $(\boldsymbol{\beta}, 0, 0)$ as in Lemma 4 in Section 2.6.2, we have

$$\begin{aligned} \begin{pmatrix} C_{n,\alpha_0,\eta}^2 n^{-1}\tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} &= \boldsymbol{\Sigma}^{-1} \begin{pmatrix} -n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) + o_p(\Delta_n) \\ o_p(\Delta_n) \\ -H(\boldsymbol{\beta}_0) + o_p(\Delta_n) \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{A}\mathbf{P}\mathbf{A}\mathbf{B}^{-1} \\ \mathbf{P}\mathbf{A}\mathbf{B}^{-1} \\ \mathbf{Q}\mathbf{A}\mathbf{B}^{-1} \end{pmatrix} \left\{ -\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) \right\} \\ &\quad + \begin{pmatrix} \mathbf{B}^{-1}\mathbf{A}\mathbf{Q}^{\intercal} \\ \mathbf{Q}^{\intercal} \\ -\mathbf{R} \end{pmatrix} \{-H(\boldsymbol{\beta}_0)\} + o_p(\Delta_n), \end{aligned}$$

where $\Delta_n = \|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\|\tilde{\boldsymbol{\gamma}}\|+\|\tilde{\boldsymbol{\nu}}\|$. Then under local alternative hypothesis $H_1 : H\{\boldsymbol{\beta}_0(t)\} = n^{-1}C_{n,\alpha_0,\eta}\mathbf{d}(t)$, we have

$$\Delta_n = \| \begin{pmatrix} \tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} \| \leq \| \begin{pmatrix} C^2_{n,\alpha_0,\eta} n^{-1}\tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} \| \leq O_p(C_{n,\alpha_0,\eta}/n) + o_p(\Delta_n),$$

which implies that $\Delta_n = O_p(C_{n,\alpha_0,\eta}/n)$.

Thus for $\tilde{\boldsymbol{\nu}}$, we have

$$\begin{aligned} \tilde{\boldsymbol{\nu}} &= -\mathbf{Q}\mathbf{A}^\intercal\mathbf{B}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + \mathbf{R}H(\boldsymbol{\beta}_0) + o_p(C_{n,\alpha_0,\eta}/n) \\ &= -\mathbf{R}C\mathbf{A}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + \mathbf{R}H(\boldsymbol{\beta}_0) + o_p(C_{n,\alpha_0,\eta}/n). \end{aligned} \tag{2.6.20}$$

Accordingly, we have $nC^{-1}_{n,\alpha_0,\eta}\mathbf{R}^{-1/2}\{\tilde{\boldsymbol{\nu}} - \mathbf{R}H(\boldsymbol{\beta}_0)\} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_q)$. Under local alternative hypothesis $H_1 : H\{\boldsymbol{\beta}_0(t)\} = n^{-1}C_{n,\alpha_0,\eta}\mathbf{d}(t)$, we have

$$nC^{-1}_{n,\alpha_0,\eta}\mathbf{R}^{-1/2}\tilde{\boldsymbol{\nu}} \xrightarrow{d} N(\mathbf{R}^{1/2}\mathbf{d}, \mathbf{I}_q).$$

Thus $2\ell(t) = \frac{n^2}{C^2_{n,\alpha_0,\eta}}\tilde{\boldsymbol{\nu}}^\intercal\mathbf{R}^{-1}\tilde{\boldsymbol{\nu}} + o_p(1) \xrightarrow{d} \chi^2_q(\mathbf{d}^\intercal\mathbf{R}\mathbf{d})$. □

2.6.2 Proofs of Propositions

In this section, we provide the proofs for all the propositions in this chapter and the existence of the RMELE $\tilde{\boldsymbol{\beta}}(t)$. An asymptotic expression for the Lagrange multiplier $\tilde{\boldsymbol{\gamma}}(t)$ in (2.3.13) is also included.

2.6.2.1 Some Useful Lemmas

We present some useful lemmas and their proofs before the proofs for the Propositions. Denote $\delta_n = \delta_{n1} + h^2$, $\delta_{n1} = (\frac{d_n \log n}{nh^2})^{\frac{1}{2}}$ where $d_n = h^2 + \bar{r}h/m$.

Lemma 1. *Under assumptions (C1)-(C3) and (C4)(i), we have*

$$\sup_{t\in[a,b]} |\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)| = O(\delta_n) \quad a.s..$$

Proof. By the expression of $\hat{\boldsymbol{\beta}}(t)$, using a Taylor expansion, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) =& \big(\mathbf{I}_p, \mathbf{0}_p\big)\{\mathbf{D}^\intercal(t)\mathbf{W}(t)\mathbf{D}(t)\}^{-1}\mathbf{D}(t)\mathbf{W}(t)\mathbf{Y} - \boldsymbol{\beta}_0(t) \\
=& \big(\mathbf{I}_p, \mathbf{0}_p\big)\big\{\sum_{i=1}^n \mathbf{D}_i^\intercal(t)\mathbf{W}_i(t)\mathbf{D}_i(t)\big\}^{-1}\big\{\sum_{i=1}^n \mathbf{D}_i^\intercal(t)\mathbf{W}_i(t)Y_i\big\} - \boldsymbol{\beta}_0(t) \\
=& \big(\mathbf{I}_p, \mathbf{0}_p\big)\big\{\sum_{i=1}^n \mathbf{D}_i^\intercal(t)\mathbf{W}_i(t)\mathbf{D}_i(t)\big\}^{-1}\big\{\sum_{i=1}^n \mathbf{D}_i^\intercal(t)\mathbf{W}_i(t)[\mathbf{B}_i(t) + \boldsymbol{\epsilon}_i]\big\},
\end{aligned}$$

where $\mathbf{B}_i(t) = \Big((t_{i1}-t)^2\mathbf{X}_{i1}^\intercal\boldsymbol{\beta}_0^{(2)}(t_{i1}^*)/2, \cdots, (t_{im_i}-t)^2\mathbf{X}_{im_i}^\intercal\boldsymbol{\beta}_0^{(2)}(t_{im_i}^*)/2\Big)^\intercal$ with t_{ij}^* between t and t_{ij} and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{12}, \ldots, \epsilon_{im_i})^\intercal$.

Observe that for denominator $\mathbf{I}(t) := \dfrac{1}{n}\sum_{i=1}^n \mathbf{D}_i^\intercal(t)\mathbf{W}_i(t)\mathbf{D}_i(t)$, we have

$$\begin{aligned}
\mathbf{I}(t) &= \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i}\begin{pmatrix} \mathbf{X}_{ij}\mathbf{X}_{ij}^\intercal K_h(t_{ij}-t) & \mathbf{X}_{ij}\mathbf{X}_{ij}^\intercal K_h(t_{ij}-t)\frac{t_{ij}-t}{h} \\ \mathbf{X}_{ij}\mathbf{X}_{ij}^\intercal K_h(t_{ij}-t)\frac{t_{ij}-t}{h} & \mathbf{X}_{ij}\mathbf{X}_{ij}^\intercal K_h(t_{ij}-t)(\frac{t_{ij}-t}{h})^2 \end{pmatrix} \\
&:= \begin{pmatrix} \mathbf{I}_{11}(t) & \mathbf{I}_{12}(t) \\ \mathbf{I}_{21}(t) & \mathbf{I}_{22}(t) \end{pmatrix}.
\end{aligned}$$

In order to get the uniform bound for $\mathbf{I}(t)$, we use Lemma 2 in [LH10] for $\mathbf{I}_{ij}(t), i, j = 1, 2$.

For $\mathbf{I}_{11}(t)$, we have

$$\begin{aligned}
\mathbb{E}\{\mathbf{I}_{11}(t)\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\intercal}(t_{ij})K_h(t_{ij}-t)\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\intercal}(t_{ij})K_h(t_{ij}-t)|t_{ij}]\right\} \\
&= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{\Gamma}(t_{ij})K_h(t_{ij}-t)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\int\mathbf{\Gamma}(s)K_h(s-t)f(s)ds = \int\mathbf{\Gamma}(s)K_h(s-t)f(s)ds \\
&= \int\mathbf{\Gamma}(t+uh)K(u)f(t+uh)du = \mathbf{\Gamma}(t)f(t)+\tilde{O}(h^2),
\end{aligned}$$

as long as $\mu_{12} < \infty$ which is true by condition (C1) and $[\mathbf{\Gamma}(t)f(t)]''$ is uniformly bounded on $t \in [a,b]$ by (C3), where $\tilde{O}$ denote uniform order for all $t \in [a,b]$ and also for the $\tilde{o}$ below. Hence, under the condition that $\mathbb{E}\left\{\sup_{t\in[a,b]}\|\mathbf{X}(t)\|^{\lambda_1}\right\} < \infty$ for some $5 \le \lambda_1 < \infty$, and $d_n^{-1}(\frac{\log n}{n})^{1-2/\lambda_1} = o(1)$, which is true under (C4)(i), by Lemma 2 in [LH10], we have

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}K_h(t_{ij}-t)-\mathbf{\Gamma}(t)f(t)\| = O(\delta_n), a.s..$$

By similar calculations for other three terms, we have

$$\mathbb{E}\{\mathbf{I}_{12}(t)\} = \int\mathbf{\Gamma}(s)K_h(s-t)\frac{s-t}{h}f(s)ds = \tilde{O}(h),$$

under $\mu_{12} < \infty$ and $[\mathbf{\Gamma}(t)f(t)]'$ uniformly bounded on $t \in [a,b]$, which are true under (C1)

and (C3) respectively. And

$$\mathbb{E}\{\mathbf{I}_{22}(t)\} = \int \mathbf{\Gamma}(s)K_h(s-t)(\frac{s-t}{h})^2 f(s)ds = \mathbf{\Gamma}(t)f(t)\mu_{12} + \tilde{O}(h^2),$$

under $[\mathbf{\Gamma}(t)f(t)]''$ is uniformly bounded on $t \in [a,b]$ by (C3). Hence in summary, we have under conditions (C1)-(C3) and (C4)(i),

$$\mathbf{I}(t) = \begin{pmatrix} \mathbf{\Gamma}(t)f(t) + \tilde{O}(\delta_n) & \tilde{O}(\delta_{n1}+h) \\ \tilde{O}(\delta_{n1}+h) & \mathbf{\Gamma}(t)f(t)\mu_{12} + \tilde{O}(\delta_n) \end{pmatrix}, \ a.s..$$

Then we have

$$\mathbf{I}^{-1}(t) = \begin{pmatrix} \mathbf{\Gamma}(t)f(t) & 0 \\ 0 & \mathbf{\Gamma}(t)f(t)\mu_{12} \end{pmatrix}^{-1} + \tilde{O}(\delta_{n1}+h), \ a.s.. \tag{2.6.21}$$

For the numerator $\mathbf{II}(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_i^{\intercal}(t)\mathbf{W}_i(t)\mathbf{B}_i(t)$, we have

$$\mathbf{II}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\begin{pmatrix} \mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}(t_{ij}-t)^2\dfrac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t) \\ \mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\dfrac{(t_{ij}-t)^3}{h}\dfrac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t) \end{pmatrix} := \begin{pmatrix} \mathbf{II}_1(t) \\ \mathbf{II}_2(t) \end{pmatrix}.$$

Similar as the denominator, under the condition that $\boldsymbol{\beta}_0(t)$ has continuous second derivative

on $t \in [a,b]$ (C3), we have

$$\begin{aligned}
\mathbb{E}\{\mathbf{II}_1(t)\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}(t_{ij}-t)^2\frac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t)\right\} \\
&= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}(\frac{t_{ij}-t}{h})^2K_h(t_{ij}-t)\right\}\tilde{O}(h^2) \\
&= \mathbf{\Gamma}(t)f(t)\mu_{12}\tilde{O}(h^2) = \tilde{O}(h^2)
\end{aligned}$$

if $\mu_{12} < \infty$ by condition (C1) and $\mathbf{\Gamma}(t)f(t)$ uniformly bounded on $t \in [a,b]$ (C3), and

$$\begin{aligned}
\mathbb{E}\{\mathbf{II}_2(t)\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\frac{(t_{ij}-t)^3}{h}\frac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t)\right\} \\
&= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}(\frac{t_{ij}-t}{h})^3K_h(t_{ij}-t)\right\}\tilde{O}(h^3) \\
&= [\mathbf{\Gamma}(t)f(t)]'\mu_{14}\tilde{O}(h^3) = \tilde{O}(h^3)
\end{aligned}$$

if $\mu_{14} < \infty$ (C1) and $[\mathbf{\Gamma}(t)f(t)]'$ uniformly bounded on $t \in [a,b]$ (C3).

By Lemma 2 in [LH10], under the condition $\mathbb{E}\left\{\sup_{t\in[a,b]}\|\mathbf{X}(t)\|^{\lambda_1}\right\} < \infty$ for some $5 \le \lambda_1 < \infty$, and $d_n^{-1}(\frac{\log n}{n})^{1-2/\lambda_1} = o(1)$ which is true under (C4)(i), we can have

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}(t_{ij}-t)^2\frac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t)\| = h^2O(\delta_{n1}+1), a.s.,$$

and

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\frac{(t_{ij}-t)^3}{h}\frac{\boldsymbol{\beta}_0^{(2)}(t_{ij}^*)}{2}K_h(t_{ij}-t)\| = h^2O(\delta_{n1}+h), a.s..$$

Note that

$$\mathbf{III}(t) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i^{\intercal}(t)\mathbf{W}_i(t)\epsilon_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\begin{pmatrix}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)\\ \mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)\frac{t_{ij}-t}{h}\end{pmatrix}.$$

Similarly, by condition (C2) and (C3), we have the following due to Lemma 2 in [LH10]

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)\| = O(\delta_{n1}), a.s.,$$

and

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)\frac{t_{ij}-t}{h}\| = O(\delta_{n1}), a.s..$$

Thus we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t) =& \left(\mathbf{I}_{p\times p}, \mathbf{0}_{p\times p}\right)\begin{pmatrix}\boldsymbol{\Gamma}(t)f(t) & 0\\ 0 & \boldsymbol{\Gamma}(t)f(t)\mu_{12}\end{pmatrix}^{-1}\\ &\times\left\{h^2\begin{pmatrix}\tilde{O}(\delta_{n1}+1)\\ \tilde{O}(\delta_{n1}+h)\end{pmatrix}+\begin{pmatrix}\tilde{O}(\delta_{n1})\\ \tilde{O}(\delta_{n1})\end{pmatrix}\right\} = h^2\tilde{O}(\delta_{n1}+1)+\tilde{O}(\delta_{n1}) = \tilde{O}(\delta_n), a.s.,\end{aligned}$$

since $\delta_n = \delta_{n1}+h^2$. □

Lemma 2. *Under conditions (C1)-(C3) and (C4)(i), we have* $\mathbb{E}(g_i\{\boldsymbol{\beta}_0(t)\}) = \tilde{O}(h^4)$ *and*

$$Var(g_i\{\boldsymbol{\beta}_0(t)\}) = \left\{\frac{1}{m_i h}\boldsymbol{\Gamma}(t)\Omega(t)f(t)\mu_{20}+\frac{m_i-1}{m_i}\boldsymbol{\Gamma}(t)\Omega(t)f^2(t)\right\}\{1+\tilde{o}(1)\}.$$

Proof. By the definition of $g_i\{\boldsymbol{\beta}_0(t)\}$, we decompose $g_i\{\boldsymbol{\beta}_0(t)\}$ as the following two parts

$$g_i\{\boldsymbol{\beta}_0(t)\} = m_i^{-1}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\{[\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)]-[\hat{\boldsymbol{\beta}}(t_{ij})-\boldsymbol{\beta}_0(t_{ij})]\}K_h(t_{ij}-t)$$
$$+ m_i^{-1}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t) := L_{1i}(t)+\xi_i(t).$$

To analyze the first term $L_{1i}(t)$ in the above expression, we further obtain the expansion for $\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)$ in the following. By the expression of $\hat{\boldsymbol{\beta}}(t)$ and a Taylor expansion, we obtain

$$\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t) = \big(\mathbf{I}_{p\times p},\mathbf{0}_{p\times p}\big)\Big\{n^{-1}\sum_{i=1}^{n}\mathbf{D}_i^{\intercal}(t)\mathbf{W}_i(t)\mathbf{D}_i(t)\Big\}^{-1}$$
$$\times\Big\{n^{-1}\sum_{i=1}^{n}\mathbf{D}_i^{\intercal}(t)\mathbf{W}_i(t)(\mathbf{B}_i(t)+\mathbf{T}_i(t)+\boldsymbol{\epsilon}_i)\Big\},$$

where $\mathbf{B}_i(t)=\frac{1}{2}(\mathbf{X}_{i1}^{\intercal}\boldsymbol{\beta}_0^{(2)}(t)(t_{i1}-t)^2,\cdots,\mathbf{X}_{im_i}^{\intercal}\boldsymbol{\beta}_0^{(2)}(t)(t_{im_i}-t)^2)^{\intercal}$ and

$$\mathbf{T}_i(t)=\frac{1}{6}(\mathbf{X}_{i1}^{\intercal}\boldsymbol{\beta}_0^{(3)}(t_{i1}^*)(t_{i1}-t)^3,\cdots,\mathbf{X}_{im_i}^{\intercal}\boldsymbol{\beta}_0^{(3)}(t_{im_i}^*)(t_{im_i}-t)^3)^{\intercal}$$

with t_{ij}^* is between t and t_{ij}. It then follows that

$$(\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t))-(\hat{\boldsymbol{\beta}}(t_{ij})-\boldsymbol{\beta}_0(t_{ij})) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\Big\{\eta_{1,kl}(t)-\eta_{1,kl}(t_{ij})$$
$$+(\eta_{2,kl}(t)-\eta_{2,kl}(t_{ij}))+(\eta_{3,kl}(t,t_1^*)-\eta_{3,kl}(t,t_2^*))\Big\}\{1+\tilde{o}_p(1)\}$$

where t_1^* is between t and t_{kl} and t_2^* is between t_{ij} and t_{kl} and

$$\begin{aligned}
\eta_{1,kl}(t) &= f^{-1}(t)\boldsymbol{\Gamma}^{-1}(t)\mathbf{X}_{kl}\epsilon_{kl}K_h(t_{kl}-t) \\
\eta_{2,kl}(t) &= \frac{1}{2}f^{-1}(t)\boldsymbol{\Gamma}^{-1}(t)\mathbf{X}_{kl}\mathbf{X}_{kl}^{\intercal}(t_{kl}-t)^2\boldsymbol{\beta}_0^{(2)}(t)K_h(t_{kl}-t) \\
\eta_{3,kl}(t,t^*) &= \frac{1}{6}f^{-1}(t)\boldsymbol{\Gamma}^{-1}(t)\mathbf{X}_{kl}\mathbf{X}_{kl}^{\intercal}(t_{kl}-t)^3\boldsymbol{\beta}_0^{(3)}(t^*)K_h(t_{kl}-t).
\end{aligned}$$

Then we can write $L_{1i}(t) = \{\boldsymbol{I}_{1i}(t) + \boldsymbol{I}_{2i}(t) + \boldsymbol{I}_{3i}(t)\}\{1+\tilde{o}_p(1)\}$ where

$$\begin{aligned}
\boldsymbol{I}_{1i}(t) &= \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\eta_{1,kl}(t)K_h(t_{ij}-t) \\
&\quad - \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\eta_{1,kl}(t_{ij})K_h(t_{ij}-t) := \boldsymbol{I}_{11,i}(t) - \boldsymbol{I}_{12,i}(t), \\
\boldsymbol{I}_{2i}(t) &= \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\eta_{2,kl}(t)K_h(t_{ij}-t) \\
&\quad - \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\eta_{2,kl}(t_{ij})K_h(t_{ij}-t) := \boldsymbol{I}_{21,i}(t) - \boldsymbol{I}_{22,i}(t) \text{ and} \\
\boldsymbol{I}_{3i}(t) &= \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{m_k}\sum_{l=1}^{m_k}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}\{\eta_{3,kl}(t,t_1^*) - \eta_{3,kl}(t_{ij},t_2^*)\}K_h(t_{ij}-t).
\end{aligned}$$

For $\boldsymbol{I}_{1i}(t)$, we have $\mathbb{E}\{\boldsymbol{I}_{1i}(t)\} = 0$ and

$$\begin{aligned}
\mathrm{Var}\{\boldsymbol{I}_{11i}(t)\} = \Big\{ &\frac{1}{n^2 m_i h^2}\sum_{k\neq i}\frac{1}{m_k}\boldsymbol{\Omega}_1(t)\Omega(t)\mu_{20}^2 \\
&+ 2\frac{m_i-1}{n^2 m_i h}\sum_{k\neq i}\frac{1}{m_k}\boldsymbol{\Omega}_1(t)\Omega(t)f(t)\mu_{20} \\
&+ \frac{m_i-1}{n^2 m_i}\sum_{k\neq i}\frac{m_k-1}{m_k}\boldsymbol{\Omega}_1(t)\Omega(t)f^2(t)\Big\}\{1+\tilde{o}(1)\}
\end{aligned}$$

where $\boldsymbol{\Omega}_1(t) = \mathbb{E}\{\mathbf{X}_i(t)\mathbf{X}_i^{\intercal}(t)\boldsymbol{\Gamma}^{-1}(t)\mathbf{X}_i(t)\mathbf{X}_i^{\intercal}(t)\}$. The leading order variance of $\boldsymbol{I}_{12i}(t)$ is

the same as that of $\boldsymbol{I}_{11i}(t)$. In summary, we have $\mathrm{Var}\{\boldsymbol{I}_{1i}(t)\} = \tilde{O}(\frac{1}{nm^2h^2})\mathbb{1}(\kappa_0 < \infty) + \tilde{O}(\frac{1}{n})\mathbb{1}(\kappa_0 = \infty)$.

By condition (C3), $\boldsymbol{\beta}_0(t)$ has continuous third derivative, and $\boldsymbol{\Gamma}(t)f(t)$, $[\boldsymbol{\Gamma}(t)f(t)]'$, $[\boldsymbol{\Gamma}(t)f(t)]''$, $\boldsymbol{\Gamma}^{-1}(t)$, $f(t)$, $f^{-1}(t)$ are uniformly bounded on $[a, b]$, we have

$$\mathbb{E}\{\boldsymbol{I}_{21,i}(t)\} = \left\{\frac{1}{2n}\boldsymbol{\Omega}_1(t) + \frac{n-1}{2n}\boldsymbol{\Gamma}(t)\right\} f(t)\boldsymbol{\beta}_0^{(2)}(t)\mu_{12}h^2 + \tilde{O}(h^4) \text{ and}$$
$$\mathbb{E}\{\boldsymbol{I}_{22,i}(t)\} = \left\{\frac{1}{2n}\boldsymbol{\Omega}_1(t) + \frac{n-1}{2n}\boldsymbol{\Gamma}(t)\right\} f(t)\boldsymbol{\beta}_0^{(2)}(t)\mu_{12}h^2 + \tilde{O}(h^4).$$

Therefore $\mathbb{E}\{\boldsymbol{I}_{2,i}(t)\} = \mathbb{E}\{\boldsymbol{I}_{21,i}(t)\} - \mathbb{E}\{\boldsymbol{I}_{22,i}(t)\} = \tilde{O}(h^4)$. To evaluate the variance of $\boldsymbol{I}_{2i}(t)$, we first evaluate the variance of $\boldsymbol{I}_{21,i}(t)$. Note that

$$\begin{aligned}
&(nm_i)^2\boldsymbol{I}_{21,i}(t)\boldsymbol{I}_{21,i}^{\intercal}(t) \\
&= \frac{1}{m_i^2}\sum_{j_1,j_2=1}^{m_i}\sum_{l_1,l_2=1}^{m_i}\mathbf{X}_{ij_1}\mathbf{X}_{ij_1}^{\intercal}\eta_{2,il_1}(t)K_h(t_{ij_i}-t)\mathbf{X}_{ij_2}\mathbf{X}_{ij_2}^{\intercal}\eta_{2,il_2}(t)K_h(t_{ij_2}-t) \\
&+ \sum_{k(\neq i)=1}^{n}\sum_{j_1,j_2=1}^{m_i}\sum_{l_1,l_2=1}^{m_k}\frac{1}{m_k^2}\mathbf{X}_{ij_1}\mathbf{X}_{ij_1}^{\intercal}\eta_{2,kl_1}(t)K_h(t_{ij_i}-t)\mathbf{X}_{ij_2}\mathbf{X}_{ij_2}^{\intercal}\eta_{2,kl_2}(t)K_h(t_{ij_2}-t) \\
&+ \sum_{(k_1\neq k_2)=1}^{n}\sum_{j_1,j_2=1}^{m_i}\sum_{l_1=1}^{m_{k_1}}\sum_{l_2=1}^{m_{k_2}}\frac{1}{m_{k_1}m_{k_2}}\mathbf{X}_{ij_1}\mathbf{X}_{ij_1}^{\intercal}\eta_{2,k_1l_1}(t)K_h(t_{ij_i}-t) \\
&\qquad\qquad \times \mathbf{X}_{ij_2}\mathbf{X}_{ij_2}^{\intercal}\eta_{2,k_2l_2}(t)K_h(t_{ij_2}-t) \\
&:= (nm_i)^2\{\mathbf{J}_1(t) + \mathbf{J}_2(t) + \mathbf{J}_3(t)\}.
\end{aligned}$$

Let $\boldsymbol{\Omega}_2(t) = \mathbb{E}\{\mathbf{X}_i(t)\mathbf{X}_i^{\intercal}(t)\mathbf{X}_i(t)\mathbf{X}_i^{\intercal}(t)\}$. It is easy to see that the dominant term of

$\mathbb{E}\{\boldsymbol{I}_{21,i}(t)\boldsymbol{I}_{21,i}^{\intercal}(t)\}$ is $\mathbb{E}\{\mathbf{J}_3(t)\}$. Careful derivation shows that, up to a scale constant,

$$\mathbb{E}\{\mathbf{J}_3(t)\} = \Big\{\frac{1}{m_i}\boldsymbol{\Omega}_2(t)f(t)\mu_{12}^2\mu_{20}h^3 + \boldsymbol{\Omega}_2(t)f^2(t)\mu_{12}^2h^4\Big\}\{1+\tilde{o}(1)\}.$$

Similar derivation shows that $\mathrm{Var}\{\boldsymbol{I}_{22,i}(t)\}$ is of the same order as $\mathrm{Var}\{\boldsymbol{I}_{21,i}(t)\}$. Therefore, in summary, we have $\mathrm{Var}\{\boldsymbol{I}_{2,i}(t)\} = \tilde{O}(h^3/m)\mathbb{1}(\kappa_0 < \infty) + \tilde{O}(h^4)\mathbb{1}(\kappa_0 = \infty)$. For $\boldsymbol{I}_{3,i}(t)$, it can be shown that $\mathbb{E}\{\boldsymbol{I}_{3,i}(t)\} = \tilde{O}(h^4)$ and

$$\mathrm{Var}\{\mathbf{I}_{3,i}(t)\} = \tilde{O}(h^6/m)\mathbb{1}(\kappa_0 < \infty) + \tilde{O}(h^7)\mathbb{1}(\kappa_0 = \infty).$$

Finally, we evaluate the order of $\xi_i(t)$. It is clear that $\mathbb{E}\{\xi_i(t)\} = 0$ and

$$\mathrm{Var}\{\xi_i(t)\} = \Big\{\frac{1}{m_i h}\boldsymbol{\Gamma}(t)\Omega(t)f(t)\mu_{20} + \frac{m_i-1}{m_i}\boldsymbol{\Gamma}(t)\Omega(t)f^2(t)\Big\}\{1+\tilde{o}(1)\}. \tag{2.6.22}$$

In summary, $\mathbb{E}\left(g_i\{\boldsymbol{\beta}_0(t)\}\right) = \tilde{O}(h^4)$ and by comparing the variance of $\xi_i(t)$ to the variances of $\boldsymbol{I}_{1,i}(t)$ to $\boldsymbol{I}_{3,i}(t)$, we have $\mathrm{Var}\left(g_i\{\boldsymbol{\beta}_0(t)\}\right) = \mathrm{Var}\{\xi_i(t)\}\{1+\tilde{o}(1)\}$. This completes the proof of this Lemma. □

Lemma 3. *Under conditions (C1)-(C4), we have for true $\boldsymbol{\beta}_0(t)$*

$$C_{n,\alpha_0,\eta}^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\} \xrightarrow{d} N(\mathbf{0}, \mathbf{B}(t)),$$

where $C_{n,\alpha_0,\eta}$ and $\mathbf{B}(t)$ are defined in Proposition 2 in Section 2.2.

Proof. Let $\xi_i(t) := m_i^{-1}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)$ and using the proof of Lemma 2,

$$g_i\{\boldsymbol{\beta}_0(t)\} = \xi_i(t)\{1+\tilde{o}_p(1)\} + \tilde{O}_p(h^4), \tag{2.6.23}$$

and $\mathbf{V}_i(t) := \mathrm{Var}\{\xi_i(t)\} = \tilde{O}\{(mh)^{-1}\}\mathbb{1}(\kappa_0 < \infty) + \tilde{O}\{1\}\mathbb{1}(\kappa_0 = \infty)$.

We will show that the asymptotic normality of $\sum_{i=1}^n g_i\{\boldsymbol{\beta}_0(t)\}$ is the same as the asymptotic normality of $\sum_{i=1}^n \xi_i(t)$.

First consider the case $\kappa_0 < \infty$, i.e. $mh \to [0, \infty)$, with (2.6.23) and condition (C4), we have

$$\frac{(mh)^{1/2}}{\sqrt{n}} \sum_{i=1}^n g_i\{\boldsymbol{\beta}_0(t)\} = \frac{(mh)^{1/2}}{\sqrt{n}} \sum_{i=1}^n \xi_i(t) + \tilde{o}_p(1). \tag{2.6.24}$$

As above, we can check that $\mathbb{E}\Big\{(mh)^{1/2} \sum_{i=1}^n \xi_i(t)/\sqrt{n}\Big\} = 0$ and

$$\begin{aligned}\mathrm{Var}\Big\{\frac{(mh)^{1/2}}{\sqrt{n}} \sum_{i=1}^n \xi_i(t)\Big\} &= \frac{1}{n}\sum_{i=1}^n [\frac{m}{m_i}\mu_{20} + \frac{m_i - 1}{m_i} f(t)]\mathbf{\Gamma}(t)\Omega(t)f(t)\{1 + \tilde{o}(1)\} \\ &\to [\bar{r}\mu_{20} + \kappa_0 f(t)]\mathbf{\Gamma}(t)\Omega(t)f(t) = \mathbf{B}(t).\end{aligned}$$

Next, we consider the case $\kappa_0 = \infty$, i.e. $mh \to \infty$. Again by (2.6.23) and condition (C4)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_i\{\boldsymbol{\beta}_0(t)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(t) + \tilde{o}_p(1). \tag{2.6.25}$$

Similarly, it can be checked that $\mathbb{E}\Big\{\sum_{i=1}^n \xi_i(t)/\sqrt{n}\Big\} = 0$ and

$$\begin{aligned}\mathrm{Var}\Big\{\sum_{i=1}^n \xi_i(t)/\sqrt{n}\Big\} &= \frac{1}{n}\sum_{i=1}^n \frac{m_i - 1}{m_i}\mathbf{\Gamma}(t)\Omega(t)f^2(t)\{1 + \tilde{o}(1)\} \\ &\to f^2(t)\mathbf{\Gamma}(t)\Omega(t) = \mathbf{B}(t).\end{aligned}$$

To show the asymptotic normality under both cases, applying the cramer-wold device, it is enough to show the asymptotic normality of $\sum_{i=1}^n \boldsymbol{\theta}^\intercal \xi_i(t)/C_{n,\alpha_0,\eta}$ for any $\boldsymbol{\theta} \in \mathbb{R}^p$ at any

fixed time point t. It remains to check the Lyapunov condition. To this end, note that

$$s_n^2 = \mathrm{Var}\{\sum_{i=1}^{n} \boldsymbol{\theta}^\mathsf{T}\xi_i(t)\} = \sum_{i=1}^{n} \boldsymbol{\theta}^\mathsf{T}\mathbf{V}_i\boldsymbol{\theta} \sim C_{n,\alpha_0,\eta}^2.$$

And on the other hand, for $m \to \infty$,

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\Big\{\Big(\boldsymbol{\theta}^\mathsf{T}\xi_i(t)\Big)^{2+\delta_0}\Big\} &= \sum_{i=1}^{n} \mathbb{E}\Big\{\Big(m_i^{-1}\sum_{j=1}^{m_i} \boldsymbol{\theta}^\mathsf{T}\mathbf{X}_{ij}\epsilon_{ij}K_h(t_{ij}-t)\Big)^{2+\delta_0}\Big\} \\ &\le C\sum_{i=1}^{n} \mathbb{E}\{\sup_t |\boldsymbol{\theta}^\mathsf{T}\mathbf{X}(t)|^{2+\delta_0}\}\mathbb{E}\{\sup_t |\epsilon(t)|^{2+\delta_0}\} \sim n\end{aligned}$$

by taking $\lambda_2 = 2+\delta_0$ in the assumption (C2). Thus we have

$$\frac{1}{s_n^{2+\delta_0}}\sum_{i=1}^{n} \mathbb{E}\Big\{\Big(\boldsymbol{\theta}^\mathsf{T}\xi_i(t)\Big)^{2+\delta_0}\Big\} \sim \frac{n}{n^{1+\delta_0/2}} \to 0, n \to \infty.$$

And similarly, for m is bounded,

$$\sum_{i=1}^{n} \mathbb{E}\Big\{\Big(\boldsymbol{\theta}^\mathsf{T}\xi_i(t)\Big)^{2+\delta_0}\Big\} \le \frac{Cn}{h^{2+\delta_0}}\mathbb{E}\{\sup_t |\boldsymbol{\theta}^\mathsf{T}\mathbf{X}(t)|^{2+\delta_0}\}\mathbb{E}\{\sup_t |\epsilon(t)|^{2+\delta_0}\} \sim n/h^{2+\delta_0}.$$

Then, it follows that

$$\frac{1}{s_n^{2+\delta_0}}\sum_{i=1}^{n} \mathbb{E}\Big\{\Big(\boldsymbol{\theta}^\mathsf{T}\xi_i(t)\Big)^{2+\delta_0}\Big\} \sim \frac{n/h^{2+\delta_0}}{(n/h)^{\frac{2+\delta_0}{2}}} = \frac{1}{n^{(\delta_0-2\alpha_0-\delta_0\alpha_0)/2}}.$$

The above ratio goes to 0 if and only if $\alpha_0 < \delta_0/2 + \delta_0$. By taking $\lambda_2 = 2+\delta_0$, this condition is equivalent to $\alpha_0 < \delta_0/2 + \delta_0 = 1 - 2/\lambda_2$. By assumption (C4), this condition is satisfied because $\alpha_0 < 1 - 2/\lambda < 1 - 2/\lambda_2$. This completes the proof of this Lemma. □

Lemma 4. *Under assumptions (C1)-(C4), and for each $t \in [a,b]$ under the null hypothesis*

$H_0 : H\{\boldsymbol{\beta}_0(t)\} = 0$, *we have*

$$2\ell(t) \xrightarrow{d} \chi^2_q.$$

Proof. First, for convenience we suppress the argument t in the functions $\boldsymbol{\beta}(t)$, $\tilde{\boldsymbol{\beta}}(t)$ and $\mathbf{A}(t)$, since we fix $t \in [a, b]$ in this proof. The proof is similar to that in [QL95].

We first obtain their derivatives with respect to the three variables $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\nu}$.

$$\frac{\partial Q_{1n}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}^\intercal} = \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal}(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta})) - g_i(\boldsymbol{\beta})\boldsymbol{\gamma}^\intercal \frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal}}{(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta}))^2},$$

$$\frac{\partial Q_{1n}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^\intercal} = -\frac{1}{n}\sum_{i=1}^{n} \frac{g_i(\boldsymbol{\beta})g_i^\intercal(\boldsymbol{\beta})}{(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta}))^2}, \quad \frac{\partial Q_{1n}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\nu}^\intercal} = 0,$$

$$\frac{\partial Q_{2n}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu})}{\partial \boldsymbol{\beta}^\intercal} = \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial^2 g_i^\intercal(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal \partial \boldsymbol{\beta}}\boldsymbol{\gamma}(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta})) - \frac{\partial g_i^\intercal(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\boldsymbol{\gamma}\boldsymbol{\gamma}^\intercal \frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal}}{(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta}))^2} + \frac{\partial C^\intercal(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal}\boldsymbol{\nu},$$

$$\frac{\partial Q_{2n}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu})}{\partial \boldsymbol{\gamma}^\intercal} = \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial g_i^\intercal(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal} - \frac{\partial g_i^\intercal(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal}\boldsymbol{\gamma} g_i^\intercal(\boldsymbol{\beta})}{(1 + \boldsymbol{\gamma}^\intercal(\boldsymbol{\beta})g_i(\boldsymbol{\beta}))^2}, \quad \frac{\partial Q_{2n}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}^\intercal} = C^\intercal(\boldsymbol{\beta}),$$

$$\frac{\partial H(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal} = C(\boldsymbol{\beta}), \quad \frac{\partial H(\boldsymbol{\beta})}{\partial \boldsymbol{\gamma}^\intercal} = 0, \quad \frac{\partial H(\boldsymbol{\beta})}{\partial \boldsymbol{\nu}^\intercal} = 0.$$

Hence, we have the following Taylor expansions of the system of equations at $(\boldsymbol{\beta}_0, 0, 0)$. Let

$\Delta_n = \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\tilde{\boldsymbol{\gamma}}\| + \|\tilde{\boldsymbol{\nu}}\|$.

$$
\begin{aligned}
0 &= Q_{1n}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\nu}}) \\
&= Q_{1n}(\boldsymbol{\beta}_0, 0, 0) + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}^\intercal}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\gamma}^\intercal}(\tilde{\boldsymbol{\gamma}} - 0) \\
&\quad + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}^\intercal}(\tilde{\boldsymbol{\nu}} - 0) + o_p(\Delta_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\intercal}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) g_i^\intercal(\boldsymbol{\beta}_0)\tilde{\boldsymbol{\gamma}} + o_p(\Delta_n), \\
0 &= Q_{2n}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\nu}}) \\
&= Q_{2n}(\boldsymbol{\beta}_0, 0, 0) + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}^\intercal}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\gamma}^\intercal}(\tilde{\boldsymbol{\gamma}} - 0) \\
&\quad + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}^\intercal}(\tilde{\boldsymbol{\nu}} - 0) + o_p(\Delta_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i^\intercal(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}} + C^\intercal(\boldsymbol{\beta}_0)\tilde{\boldsymbol{\nu}} + o_p(\Delta_n),
\end{aligned}
$$

and $0 = H(\tilde{\boldsymbol{\beta}}) = H(\boldsymbol{\beta}_0) + C(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(\Delta_n) = C(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(\Delta_n)$. Putting the above equations into a matrix form, we obtain

$$
\begin{pmatrix}
-n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) + o_p(\Delta_n) \\
o_p(\Delta_n) \\
o_p(\Delta_n)
\end{pmatrix}
= \boldsymbol{\Sigma}_n
\begin{pmatrix}
C_{n,\alpha_0,\eta}^2 n^{-1}\tilde{\boldsymbol{\gamma}} \\
\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\
\tilde{\boldsymbol{\nu}}
\end{pmatrix}.
$$

where

$$
\boldsymbol{\Sigma}_n =
\begin{pmatrix}
-C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) g_i^\intercal(\boldsymbol{\beta}_0) & n^{-1}\sum_{i=1}^{n} \frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\intercal} & 0 \\
n^{-1}\sum_{i=1}^{n} \frac{\partial g_i^\intercal(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} & 0 & C^\intercal(\boldsymbol{\beta}_0) \\
0 & C(\boldsymbol{\beta}_0) & 0
\end{pmatrix}.
$$

Then we have

$$\boldsymbol{\Sigma}_n \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma} = \begin{pmatrix} -\mathbf{B} & \mathbf{A} & 0 \\ \mathbf{A} & 0 & \mathbf{C}^\intercal \\ 0 & \mathbf{C} & 0 \end{pmatrix}.$$

By calculation, we have

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} -\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{A}\mathbf{P}\mathbf{A}\mathbf{B}^{-1} & \mathbf{B}^{-1}\mathbf{A}\mathbf{P} & \mathbf{B}^{-1}\mathbf{A}\mathbf{Q}^\intercal \\ \mathbf{P}\mathbf{A}\mathbf{B}^{-1} & \mathbf{P} & \mathbf{Q}^\intercal \\ \mathbf{Q}\mathbf{A}\mathbf{B}^{-1} & \mathbf{Q} & -\mathbf{R} \end{pmatrix},$$

where $\mathbf{P} = \mathbf{V}(\mathbf{I} - \mathbf{C}^\intercal\mathbf{Q})$, $\mathbf{R} = (\mathbf{C}\mathbf{V}\mathbf{C}^\intercal)^{-1}$, $\mathbf{Q} = \mathbf{R}\mathbf{C}\mathbf{V}$, $\mathbf{V} = (\mathbf{A}\mathbf{B}^{-1}\mathbf{A})^{-1}$. Thus we have the following

$$\begin{pmatrix} C_{n,\alpha_0,\eta}^2 n^{-1}\tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} = \boldsymbol{\Sigma}^{-1} \begin{pmatrix} -n^{-1}\sum_{i=1}^n g_i(\boldsymbol{\beta}_0) \\ 0 \\ 0 \end{pmatrix} + o_p(\Delta_n)$$

By this, we could figure out that

$$\begin{aligned} \Delta_n = \| \begin{pmatrix} \tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} \| &\leq \| \begin{pmatrix} C_{n,\alpha_0,\eta}^2 n^{-1}\tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} \| \\ &= \|\boldsymbol{\Sigma}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \left\{ -\frac{1}{n}\sum_{i=1}^n g_i(\boldsymbol{\beta}_0) \right\} + o_p(\Delta_n)\| \leq O_p(C_{n,\alpha_0,\eta}/n) + o_p(\Delta_n), \end{aligned}$$

which implies that $\Delta_n = O_p(C_{n,\alpha_0,\eta}/n)$.

In summary of the above results, we have

$$\begin{pmatrix} C^2_{n,\alpha_0,\eta} n^{-1}\tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\nu}} \end{pmatrix} = \begin{pmatrix} -\mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{APAB}^{-1} \\ \mathbf{PAB}^{-1} \\ \mathbf{QAB}^{-1} \end{pmatrix} \left\{ -\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0) \right\} + o_p(C_{n,\alpha_0,\eta}/n). \tag{2.6.26}$$

Thus we have the asymptotic expression for $\tilde{\boldsymbol{\nu}}$,

$$\tilde{\boldsymbol{\nu}} = -\mathbf{RCA}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + o_p(C_{n,\alpha_0,\eta}/n). \tag{2.6.27}$$

For the asymptotic expression of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, (2.6.26) together with (2.6.27) gives

$$\begin{aligned} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= [-\mathbf{A}^{-1} + \mathbf{VC}^{\intercal}\mathbf{RCA}^{-1}]\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + o_p(C_{n,\alpha_0,\eta}/n) \\ &= -\mathbf{A}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + \mathbf{VC}^{\intercal}\mathbf{RCA}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} + o_p(C_{n,\alpha_0,\eta}/n) \\ &= -\mathbf{A}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}_0)\} - \mathbf{VC}^{\intercal}\tilde{\boldsymbol{\nu}} + o_p(C_{n,\alpha_0,\eta}/n). \end{aligned} \tag{2.6.28}$$

Using the expression of $\boldsymbol{\gamma}$ in (2.6.40) and the above asymptotic expression for $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$,

the empirical log-likelihood ratio statistic can be written as

$$
\begin{aligned}
2\ell(t) &= 2\sum_{i=1}^{n}\tilde{\boldsymbol{\gamma}}^{\intercal}g_i(\tilde{\boldsymbol{\beta}}) - \sum_{i=1}^{n}\tilde{\boldsymbol{\gamma}}^{\intercal}g_i(\tilde{\boldsymbol{\beta}})g_i^{\intercal}(\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\gamma}} + o_p(1) \\
&= n(\frac{1}{n}\sum_{i=1}^{n}g_i^{\intercal}(\tilde{\boldsymbol{\beta}}))\frac{n}{C_{n,\alpha_0,\eta}^2}\mathbf{B}^{-1}(\frac{1}{n}\sum_{i=1}^{n}g_i(\tilde{\boldsymbol{\beta}})) + o_p(1) \\
&= \frac{n^2}{C_{n,\alpha_0,\eta}^2}\tilde{\boldsymbol{\nu}}^{\intercal}\mathbf{CVAB}^{-1}\mathbf{AVC}^{\intercal}\tilde{\boldsymbol{\nu}} + o_p(1) = \frac{n^2}{C_{n,\alpha_0,\eta}^2}\tilde{\boldsymbol{\nu}}^{\intercal}\mathbf{R}^{-1}\tilde{\boldsymbol{\nu}} + o_p(1).
\end{aligned}
$$

By (2.6.27), we have

$$
2\ell(t) = \frac{1}{C_{n,\alpha_0,\eta}^2}\{\sum_{i=1}^{n}g_i(\boldsymbol{\beta}_0)\}^{\intercal}\mathbf{A}^{-1}\mathbf{C}^{\intercal}\mathbf{RCA}^{-1}\{\sum_{i=1}^{n}g_i(\boldsymbol{\beta}_0)\} + o_p(1). \tag{2.6.29}
$$

We see that $\mathbb{E}(\mathbf{R}^{1/2}\mathbf{CA}^{-1}\{\sum_{i=1}^{n}g_i(\boldsymbol{\beta}_0)) = 0$ and as $n \to \infty$,

$$
\begin{aligned}
&C_{n,\alpha_0,\eta}^{-1}\mathrm{Var}\Big(\mathbf{R}^{1/2}\mathbf{CA}^{-1}\{\sum_{i=1}^{n}g_i(\boldsymbol{\beta}_0)\}\Big) \to \mathbf{R}^{1/2}\mathbf{CA}^{-1}\mathbf{BA}^{-1}\mathbf{C}^{\intercal}\mathbf{R}^{1/2} \\
&= \mathbf{R}^{1/2}\mathbf{C}\{\mathbf{ABA}\}^{-1}\mathbf{C}^{\intercal}\mathbf{R}^{1/2} = \mathbf{R}^{1/2}\mathbf{CVC}^{\intercal}\mathbf{R}^{1/2} = \mathbf{R}^{1/2}\mathbf{R}^{-1}\mathbf{R}^{1/2} = \mathbf{I}_{q\times q}
\end{aligned}
$$

Thus, by central limit theorem, we have $\mathbf{R}^{1/2}\mathbf{CA}^{-1}\{C_{n,\alpha_0,\eta}^{-1}\sum_{i=1}^{n}g_i(\boldsymbol{\beta}_0)\} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_q)$. Then by (2.6.29), we have $2\ell(t) \xrightarrow{d} \chi_q^2$. □

Denote $\delta_n^* = (\frac{d_n \log n}{nh^2})^{\frac{1}{2}-\kappa} + h^{2-\kappa}$ for some $0 < \kappa < \frac{1}{6}$.

Lemma 5. *Under assumptions (C1)-(C3) and (C4)(i), we have the solution $\check{\boldsymbol{\beta}}(t)$ to the estimating equation (2.2.7) satisfies*

(a) $\sup_{t\in[a,b]}\|\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\| = O(\delta_{n1} + h^4), a.s.$.

(b) And for each $t \in [a,b]$, in the sphere $\Big\{\boldsymbol{\beta}(t) : \sup_{t\in[a,b]}\|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)\| \le \delta_n^*\Big\}$, where

$\boldsymbol{\beta}_0(t)$ *is the true parameter, we have*

$$2\ell(t) = n^2 C_{n,\alpha_0,\eta}^{-2} H^{\intercal}\{\check{\boldsymbol{\beta}}(t)\}\mathbf{R}(t)H\{\check{\boldsymbol{\beta}}(t)\} + o_p(nh^4/C_{n,\alpha_0,\eta}).$$

Proof. We first prove (a). Using the estimating equation (2.3), one obtain

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^{n} g_i\{\check{\boldsymbol{\beta}}(t)\} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\epsilon_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t) \\ + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\Delta_{\check{\beta},ij}^{\intercal}(t)\mathbf{X}_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t), \end{aligned}$$

where $\Delta_{\check{\beta},ij}(t) = [\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] - [\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] - [\hat{\boldsymbol{\beta}}(t_{ij}) - \boldsymbol{\beta}_0(t_{ij})]$

It follows that

$$\begin{aligned} &\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}^{\intercal}\mathbf{X}_{ij}K_h(t_{ij}-t)\Big\}[\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\epsilon_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\Big\{[\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] - [\hat{\boldsymbol{\beta}}(t_{ij}) - \boldsymbol{\beta}_0(t_{ij})]\Big\}^{\intercal}\mathbf{X}_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t) = \bar{g}_n\{\boldsymbol{\beta}_0(t)\}, \end{aligned} \tag{2.6.30}$$

Since we have $\bar{g}_n\{\boldsymbol{\beta}_0(t)\} = \tilde{O}_p(\delta_{n1} + h^4)$, and we also know that from the proof of Lemma 1,

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\intercal}K_h(t_{ij}-t) - \boldsymbol{\Gamma}(t)f(t)\| = O(\delta_n), a.s..$$

Thus (2.6.30) gives $\sup_{t\in[a,b]}\|\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\| = O(\delta_{n1} + h^4), a.s.$. This completes the proof of part (a).

For (b), we have the following Taylor expansion for $\frac{1}{n}\sum_{i=1}^{n} g_i\{\check{\boldsymbol{\beta}}(t)\}$ by (a) for each $t \in [a,b]$, we have $\|\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\| = O_p(C_{n,\alpha_0,\eta}/n + h^4)$

$$
\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} g_i\{\check{\boldsymbol{\beta}}(t)\} = \frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\} + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i\{\boldsymbol{\beta}_0(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)}[\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] + o_p(C_{n,\alpha_0,\eta}/n + h^4) \\
&= \frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\} + \mathbf{A}(t)[\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] + o_p(C_{n,\alpha_0,\eta}/n + h^4),
\end{aligned}
\tag{2.6.31}
$$

which gives

$$
\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) = -\mathbf{A}^{-1}(t)\{\frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\}\} + o_p(C_{n,\alpha_0,\eta}/n + h^4). \tag{2.6.32}
$$

The Taylor expansion for $H\{\check{\boldsymbol{\beta}}(t)\}$ around $\boldsymbol{\beta}_0(t)$ can be expressed as follows by plugging in (2.6.32)

$$
\begin{aligned}
H\{\check{\boldsymbol{\beta}}(t)\} &= H\{\boldsymbol{\beta}_0(t)\} + \mathbf{C}(t)[\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] + o_p(C_{n,\alpha_0,\eta}/n + h^4) \\
&= H\{\boldsymbol{\beta}_0(t)\} - \mathbf{C}(t)\mathbf{A}^{-1}(t)\{\frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\}\} + o_p(C_{n,\alpha_0,\eta}/n + h^4) \\
&= H\{\boldsymbol{\beta}_0(t)\} + \mathbf{R}^{-1}(t)\tilde{\boldsymbol{\nu}}(t) - H\{\boldsymbol{\beta}_0(t)\} + o_p(C_{n,\alpha_0,\eta}/n + h^4) \\
&= \mathbf{R}^{-1}(t)\tilde{\boldsymbol{\nu}}(t) + o_p(C_{n,\alpha_0,\eta}/n + h^4),
\end{aligned}
\tag{2.6.33}
$$

where the second-to-last equality is due to similar result as (2.6.27) for general $H\{\boldsymbol{\beta}_0(t)\}$.

Thus we could easily see from the proof of Lemma 4 that

$$
\begin{aligned}
2\ell(t) &= \frac{n^2}{C_{n,\alpha_0,\eta}^2}\tilde{\boldsymbol{\nu}}^{\intercal}\mathbf{R}^{-1}\tilde{\boldsymbol{\nu}} + o_p(nh^4/C_{n,\alpha_0,\eta}) \\
&= \frac{n^2}{C_{n,\alpha_0,\eta}^2}H^{\intercal}\{\check{\boldsymbol{\beta}}(t)\}\mathbf{R}(t)H\{\check{\boldsymbol{\beta}}(t)\} + o_p(nh^4/C_{n,\alpha_0,\eta}).
\end{aligned}
$$

□

2.6.2.2 Proof of Propositions

In this section, we provide the proof for the Propositions in this chapter.

Proof of Proposition 1. By (2.6.32), we have

$$
\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) = -\mathbf{A}^{-1}(t)\{\frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\}\} + o_p(C_{n,\alpha_0,\eta}/n + h^4).
$$

And by Lemma 2, we have

$$
g_i\{\boldsymbol{\beta}_0(t)\} = \xi_i(t)\{1 + \tilde{o}_p(1)\} + \tilde{O}_p(h^4).
$$

Combining these two results together, we have $\check{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t) = -\mathbf{A}^{-1}(t)\bar{\boldsymbol{\xi}}_n(t)\{1 + \tilde{o}_p(1)\} + \tilde{O}_p(h^4)$.

And for $\mathrm{Var}\{\bar{\boldsymbol{\xi}}_n(t)\}$, from (2.6.22) in the proof of Lemma 2, we can easily get (2.2.9) defined in the proposition.

□

Proof of Proposition 2. By Lemma 3, and Proposition 1, and under the bandwidth condition

$(C4)$ which makes the bias negligible, we have

$$nC_{n,\alpha_0,\eta}^{-1}\left\{\check{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)\right\}\xrightarrow{d}N(\mathbf{0},\mathbf{V}(t))$$

where $\mathbf{V}(t)=\mathbf{A}^{-1}(t)\mathbf{B}(t)\mathbf{A}^{-1}(t)$. □

Proof of Proposition 3. By (b) of Lemma 5, we have

$$2\ell(t)=\frac{n^2}{C_{n,\alpha_0,\eta}^2}H^{\intercal}\{\check{\boldsymbol{\beta}}(t)\}\mathbf{R}(t)H\{\check{\boldsymbol{\beta}}(t)\}+o_p(nh^4/C_{n,\alpha_0,\eta}).$$

From (2.6.33), we have that under $H_0: H\{\boldsymbol{\beta}_0(t)\}=0$,

$$\begin{aligned}\mathbf{R}^{1/2}(t)H\{\check{\boldsymbol{\beta}}(t)\}&=-\mathbf{R}^{1/2}(t)\mathbf{C}(t)\mathbf{A}^{-1}(t)\{\frac{1}{n}\sum_{i=1}^{n}g_i\{\boldsymbol{\beta}_0(t)\}\}\{1+\tilde{o}_p(1)\}\\&=-\mathbf{R}^{1/2}(t)\mathbf{C}(t)\mathbf{A}^{-1}(t)\bar{\boldsymbol{\xi}}_n(t)\{1+\tilde{o}_p(1)\}+\tilde{O}_p(h^4)\\&=-\mathbf{G}(t)\bar{\boldsymbol{\xi}}_n(t)\{1+\tilde{o}_p(1)\}+\tilde{O}_p(h^4).\end{aligned}$$

By $\mathbf{U}_n(t)=nC_{n,\alpha_0,\eta}^{-1}\mathbf{G}(t)\bar{\boldsymbol{\xi}}_n(t)$, we have

$$2\ell(t)=\mathbf{U}_n(t)^{\intercal}\mathbf{U}_n(t)+O_p(nh^4/C_{n,\alpha_0,\eta}).$$

□

2.6.2.3 Existence of RMELE and the asymptotic expression for $\tilde{\gamma}$

In this section, we study the existence of RMELE $\tilde{\boldsymbol{\beta}}(t)$ and the order of the Lagrange multiplier $\tilde{\boldsymbol{\gamma}}(t)$. To this end, define $\delta_n^*=(\frac{d_n\log n}{nh^2})^{\frac{1}{2}-\kappa}+h^{2-\kappa}$ for some $0<\kappa<\frac{1}{6}$ where

$d_n = h^2 + \bar{r}h/m$.

Lemma 6. *Under assumptions (C1)-(C3) and (C4)(i), in the sphere*

$$\left\{ \boldsymbol{\beta}(t) : \sup_{t\in[a,b]} \|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)\| \leq \delta_n^* \right\}, \tag{2.6.34}$$

where $\boldsymbol{\beta}_0(t)$ is the true parameter, we have (a) $\sup_t \|n^{-1}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\}\| = O_p(\delta_n^*)$; (b) $\sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| = o_p(\delta_n^{'-1})$ with $\delta_n' = n\delta_n^*/C_{n,\alpha_0,\eta}^2 \leq \delta_n^*$; and (c)

$$\lim_{n\to\infty} \mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\} g_i^\intercal\{\boldsymbol{\beta}(t)\} > 0) = 1.$$

Proof. For (a), notice that $\frac{1}{n}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\} = \mathbf{T}_1(t) + \mathbf{T}_2(t)$ where

$$\mathbf{T}_1(t) = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} \epsilon_{ij}\mathbf{X}_{ij}K_h(t_{ij} - t)$$

and

$$\mathbf{T}_2(t) = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} \Delta_{\beta,ij}^\intercal(t)\mathbf{X}_{ij}\mathbf{X}_{ij}K_h(t_{ij} - t)$$

where $\Delta_{\beta,ij}(t) = [\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)] - [\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] - [\hat{\boldsymbol{\beta}}(t_{ij}) - \boldsymbol{\beta}_0(t_{ij})]$.

For $\mathbf{T}_1(t)$, by Lemma 1 in [LH10] for the process $\epsilon(t)\mathbf{X}(t)$, under the condition (C2), as we proved in Lemma 1, we have since $\mathbb{E}\{\mathbf{T}_1(t)\} = 0$ and hence $\sup_t \|\mathbf{T}_1(t)\| = O(\delta_{n1}^*), a.s.$.

For $\mathbf{T}_2(t)$, by Lemma 1 and the assumption for $\boldsymbol{\beta}(t)$ in (2.6.34),

$$\begin{aligned}\sup_t \|\mathbf{T}_2(t)\| \leq & \sup_t \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\|\Delta_{\beta,ij}(t)\|\|\mathbf{X}_{ij}\|^2 K_h(t_{ij}-t) \\ \leq & (2\sup_t\|\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)\| + \sup_t\|\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)\|) \\ & \times \sup_t \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\|\mathbf{X}_{ij}\|^2 K_h(t_{ij}-t) = O_p(\delta_n^*).\end{aligned}$$

Thus we have $\sup_t \|n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}\| = O_p(\delta_n^*)$. This finishes the proof for part (a).

For proving part (b), note that,

$$\begin{aligned}&\sup_t \|g_i\{\boldsymbol{\beta}(t)\}\| \\ &\leq \sup_t \|\frac{1}{m_i}\sum_{j=1}^{m_i}\epsilon_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t)\| + \sup_t\|\frac{1}{m_i}\sum_{j=1}^{m_i}\Delta_{\beta,ij}^{\intercal}(t)\mathbf{X}_{ij}\mathbf{X}_{ij}K_h(t_{ij}-t)\| \\ &\leq \sup_t\|\epsilon_i(t)\mathbf{X}_i(t)\|\sup_t\frac{1}{m_i}\sum_{j=1}^{m_i}K_h(t_{ij}-t) \\ &+ \left\{2\sup_t\|\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)\| + \sup_t\|\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)\|\right\}\sup_t\|\mathbf{X}_i(t)\|^2\sup_t\frac{1}{m_i}\sum_{j=1}^{m_i}K_h(t_{ij}-t) \\ &\leq \Big(\sup_t\|\epsilon_i(t)\mathbf{X}_i(t)\| + C_1\delta_n^*\sup_t\|\mathbf{X}_i(t)\|^2\Big)\sup_t\frac{1}{m_i}\sum_{j=1}^{m_i}K_h(t_{ij}-t).\end{aligned}$$

If m_i's are bounded, then we have $\sup_t {m_i}^{-1}\sum_{j=1}^{m_i}K_h(t_{ij}-t) = O_p(1/h)$. And if m_i's tend to infinity, then by the theorem in [Sil78] we have $\sup_t {m_i}^{-1}\sum_{j=1}^{m_i}K_h(t_{ij}-t) = O_p(1)$ under the regularity conditions of the kernel function in (C1).

For the case m_i's bounded, we have $\delta_n' = h\delta_n^*$ and

$$\begin{aligned}\sup_t \|g_i\{\boldsymbol{\beta}(t)\}\| &\le \sup_t \frac{1}{m_i}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\{\sup_t \|\epsilon_i(t)\mathbf{X}_i(t)\| + C_1\delta_n^* \sup_t \|\mathbf{X}_i(t)\|^2\} \\ &\le \frac{C}{h}(\sup_t \|\epsilon_i(t)\mathbf{X}_i(t)\| + \delta_n^* \sup_t \|\mathbf{X}_i(t)\|^2).\end{aligned}$$

Then we have, for any $\epsilon > 0$, by assumption (C4),

$$\begin{aligned}&\mathbb{P}(\max_{1\le i\le n}\sup_t \|g_i\{\boldsymbol{\beta}(t)\}\| > \frac{\epsilon}{\delta_n'}) \\ \le& n\mathbb{P}\left\{\frac{C}{h}(\sup_t \|\epsilon_i(t)\mathbf{X}_i(t)\| + \delta_n^* \sup_t \|\mathbf{X}_i(t)\|^2) > \frac{\epsilon}{h\delta_n^*}\right\} \\ \le& n\mathbb{P}(\sup_t \|\epsilon(t)\mathbf{X}(t)\| > \frac{\epsilon}{2C\delta_n^*}) + n\mathbb{P}(\sup_t \|\mathbf{X}_i(t)\|^2 > \frac{\epsilon}{2C\delta_n^{*2}}) \\ \le& n\mathbb{E}\{[\sup_t \|\epsilon(t)\mathbf{X}(t)\|]^\lambda\}(\frac{2C\delta_n^*}{\epsilon})^\lambda + n\mathbb{E}\{[\sup_t \|\mathbf{X}(t)\|]^{\lambda_1}\}(\frac{2C\delta_n^{*2}}{\epsilon})^{\lambda_1/2} \\ \le& n\Big((\frac{2C\delta_n^*}{\epsilon})^\lambda \mathbb{E}\{\sup_t \|\epsilon(t)\|^\lambda\} + (\frac{2C\delta_n^{*2}}{\epsilon})^{\lambda_1/2}\Big)\mathbb{E}\{[\sup_t \|\mathbf{X}(t)\|]^{\lambda_1}\} \\ \le& Cn\{(\delta_n^*)^\lambda + (\delta_n^*)^{\lambda_1}\} \le Cn(\delta_n^*)^\lambda \to 0,\end{aligned}$$

where $\lambda = \min\{\lambda_1, \lambda_2\}$. This implies $\sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| = o_p(\delta_n^{'-1})$.

For the case that m_i's tend to infinity,we have

$$\begin{aligned}\sup_t \|g_i\{\boldsymbol{\beta}(t)\}\| &\le \sup_t \frac{1}{m_i}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\{\sup_t \|\epsilon_i(t)\mathbf{X}_i(t)\| + C_1\delta_n^* \sup_t \|\mathbf{X}_i(t)\|^2\} \\ &\le C\left\{\sup_t \|\epsilon_i(t)\mathbf{X}_i(t)\| + \delta_n^* \sup_t \|\mathbf{X}_i(t)\|^2\right\}.\end{aligned}$$

Then we have, for any $\epsilon > 0$, by assumption (C4),

$$\mathbb{P}\left\{\max_{1\le i\le n}\sup_t \|g_i\{\boldsymbol{\beta}(t)\}\| > \frac{\epsilon}{\delta_n^*}\right\} \le Cn\left\{(\delta_n^*)^\lambda + (\delta_n^*)^{\lambda_1}\right\} \le Cn(\delta_n^*)^\lambda \to 0,$$

where $\lambda = \min\{\lambda_1, \lambda_2\}$. This implies $\sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| = o_p(\delta_n^{*-1}) = o_p(\delta_n^{'-1})$. This completes the proof of part (b).

For (c), we need to show that, for any $\mathbf{u} \in \mathbb{R}^p$,

$$\lim_{n\to\infty}\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \mathbf{u}^\intercal g_i(\boldsymbol{\beta}(t))g_i^\intercal(\boldsymbol{\beta}(t))\mathbf{u} > 0) = 1. \tag{2.6.35}$$

In fact, note that

$$\begin{aligned}
&C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\}g_i^\intercal\{\boldsymbol{\beta}(t)\}\\
&= C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\\
&\quad + C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\Delta_{\beta,ij}^\intercal(t)\mathbf{X}_{ij}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal\epsilon_{il}K_h(t_{ij}-t)K_h(t_{il}-t)\\
&\quad + C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\Delta_{\beta,ij}^\intercal(t)\mathbf{X}_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal\epsilon_{ij}K_h(t_{ij}-t)K_h(t_{il}-t)\\
&\quad + C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\Delta_{\beta,ij}^\intercal(t)\mathbf{X}_{ij}\Delta_{\beta,il}^\intercal(t)\mathbf{X}_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\\
&= C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^n \frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t) + \tilde{o}_p(1),
\end{aligned}$$

by Lemma 1, and the assumption (2.6.34) for $\boldsymbol{\beta}(t)$. Thus we have for any $\epsilon_u > 0$,

$$
\begin{aligned}
&\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \mathbf{u}^\intercal g_i\{\boldsymbol{\beta}(t)\} g_i^\intercal\{\boldsymbol{\beta}(t)\}\mathbf{u} > 0) \\
\geq &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \mathbf{u}^\intercal g_i\{\boldsymbol{\beta}(t)\} g_i^\intercal\{\boldsymbol{\beta}(t)\}\mathbf{u} > \epsilon_u) \\
= &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} + \tilde{o}_p(1) > \epsilon_u) \\
\geq &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} > 2\epsilon_u, |o_p(1)| < \epsilon_u) \\
= &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} > 2\epsilon_u) \\
- &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} > 2\epsilon_u, |o_p(1)| \geq \epsilon_u) \\
\geq &\mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} > 2\epsilon_u) - \mathbb{P}(|o_p(1)| \geq \epsilon_u).
\end{aligned}
$$

Now since $\lim_{n\to\infty} \mathbb{P}(|o_p(1)| \geq \epsilon_u) = 0$, for proving (2.6.35) we only need to prove that for some $\epsilon_u > 0$,

$$
\lim_{n\to\infty} \mathbb{P}(\inf_t C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^n \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \mathbf{u}^\intercal \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u} > \epsilon_u) = 1. \quad (2.6.36)
$$

To this end, note that

$$
\begin{aligned}
&\mathbb{E}\left\{ \frac{1}{m_i^2} \sum_{j,l=1}^{m_i} \epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^\intercal K_h(t_{ij}-t)K_h(t_{il}-t) \right\} = \mathrm{Var}(\xi_i(t)) \\
&= \tilde{O}\{(mh)^{-1}\}\mathbb{1}\{\kappa_0 = 0\} + \tilde{O}(1)\mathbb{1}\{\kappa_0 = \infty\}.
\end{aligned}
$$

By the strong law of large numbers, we have

$$C_{n,\alpha_0,\eta}^{-2}\sum_{i=1}^{n}\frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^{\intercal}K_h(t_{ij}-t)K_h(t_{il}-t)\to\mathbf{L}(t),a.s.,$$

where $\mathbf{L}(t)=\bar{r}\mathbf{\Gamma}(t)\Omega(t)f(t)\mu_{20}\mathbb{1}\{\kappa_0=0\}+\mathbf{\Gamma}(t)\Omega(t)f^2(t)\mathbb{1}\{\kappa_0=\infty\}$. By taking $\epsilon_u=\frac{1}{2}\inf_t\mathbf{u}^{\intercal}\mathbf{L}(t)\mathbf{u}>0$, we have

$$\lim_{n\to\infty}\mathbb{P}\Big\{\inf_t\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i^2}\sum_{j,l=1}^{m_i}\mathbf{u}^{\intercal}\epsilon_{ij}\epsilon_{il}\mathbf{X}_{ij}\mathbf{X}_{il}^{\intercal}K_h(t_{ij}-t)K_h(t_{il}-t)\mathbf{u}>\epsilon_u\Big\}=1.$$

Hence (c) is proved. □

Lemma 7. *Under assumptions (C1)-(C3) and (C4)(i), in the sphere*

$$\left\{\boldsymbol{\beta}(t):\sup_{t\in[a,b]}\|\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)\|\le\delta_n^*\right\},\tag{2.6.37}$$

where $\boldsymbol{\beta}_0(t)$ is the true parameter, the equation $Q_{1n}\{\boldsymbol{\beta}(t),\boldsymbol{\gamma}(t)\}=0$ almost surely has root $\boldsymbol{\gamma}(t)=\boldsymbol{\gamma}\{\boldsymbol{\beta}(t)\}$ and $\sup_{t\in[a,b]}\|\boldsymbol{\gamma}(t)\|=O_p(\delta_n')$,where $\delta_n'=n\delta_n^*/C_{n,\alpha_0,\eta}^2\le\delta_n^*$.

Proof. Similar to the proof in [Owe90], let $\boldsymbol{\gamma}(t):=\rho(t)\boldsymbol{\theta}(t)$ with $\|\boldsymbol{\theta}(t)\|=1$ and $\rho(t)\ge0$, and then from the equation

$$Q_{1n}\{\boldsymbol{\beta}(t),\boldsymbol{\gamma}(t)\}=\frac{1}{n}\sum_i\frac{g_i\{\boldsymbol{\beta}(t)\}}{1+\boldsymbol{\gamma}^{\intercal}(t)g_i\{\boldsymbol{\beta}(t)\}}=0,$$

we have

$$\frac{\rho(t)\boldsymbol{\theta}^{\intercal}(t)\mathbf{S}(t)\boldsymbol{\theta}(t)}{1+\rho(t)\sup_t\max_i\|g_i\{\boldsymbol{\beta}(t)\}\|}-\frac{1}{n}|\boldsymbol{\theta}^{\intercal}(t)\sum_{i=1}^{n}g_i\{\boldsymbol{\beta}(t)\}|\le0,\tag{2.6.38}$$

where $\mathbf{S}(t)=\frac{1}{n}\sum_{i=1}^{n}g_i\{\boldsymbol{\beta}(t)\}g_i^{\intercal}\{\boldsymbol{\beta}(t)\}$. By applying (a)-(c) in Lemma 6 and (2.6.38), we

have

$$\rho(t)\boldsymbol{\theta}^{\mathsf{T}}(t)nC_{n,\alpha_0,\eta}^{-2}\mathbf{S}(t)\boldsymbol{\theta}(t) \leq \{1+\rho(t)\sup_t \max_i \|g_i(\boldsymbol{\beta}(t))\|\}\frac{1}{C_{n,\alpha_0,\eta}^2}|\boldsymbol{\theta}^{\mathsf{T}}(t)\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\}|$$
$$= \{1+\rho(t)\tilde{o}_p(\delta_n^{'-1})\}\tilde{O}_p(\delta_n') = \tilde{O}_p(\delta_n') + \rho(t)\tilde{o}_p(1),$$

which implies

$$\rho(t) \leq \frac{\tilde{O}_p(\delta_n')}{\boldsymbol{\theta}^{\mathsf{T}}(t)nC_{n,\alpha_0,\eta}^{-2}\mathbf{S}(t)\boldsymbol{\theta}(t) + o_p(1)} \sim \tilde{O}_p(\delta_n'),$$

since $\mathbf{S}(t) \sim C_{n,\alpha_0,\eta}^2/n$ uniformly for $t \in [a,b]$. Namely we proved $\sup_t \|\boldsymbol{\gamma}(t)\| = O_p(\delta_n')$. □

Remark 4. *For $\boldsymbol{\gamma}\{\boldsymbol{\beta}_0(t)\}$, we have $\sup_t \|\boldsymbol{\gamma}(\boldsymbol{\beta}_0(t))\| = O_p(n\delta_n/C_{n,\alpha_0,\eta}^2)$. This is because Lemma 1, Lemma 6 and Lemma 7 are still true if we replace $\boldsymbol{\beta}(t)$ by $\boldsymbol{\beta}_0(t)$ and δ_n^* by δ_n. This implies that $\sup_t \|\boldsymbol{\gamma}(\boldsymbol{\beta}_0(t))\| = O_p\{(\log(n)/nh)^{1/2}\}$ for sparse data and $\sup_t \|\boldsymbol{\gamma}(\boldsymbol{\beta}_0(t))\| = O_p\{(\log(n)/n)^{1/2}\}$ for dense data.*

Expression for $\boldsymbol{\gamma}(t)$: From the equation Q_{1n}, we have

$$\begin{aligned} 0 = Q_{1n}\{\boldsymbol{\beta}(t),\boldsymbol{\gamma}(t)\} &= n^{-1}\sum_{i=1}^n \frac{g_i\{\boldsymbol{\beta}(t)\}}{1+\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}} \\ &= n^{-1}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\} - n^{-1}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\}g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}\boldsymbol{\gamma}(t) \\ &\quad + n^{-1}\sum_{i=1}^n g_i\{\boldsymbol{\beta}(t)\}\frac{[\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}]^2}{1+\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}} \end{aligned} \tag{2.6.39}$$

In the following, we want to show that the order of the third term is $\tilde{o}_p(\delta_n')$. To this end, we firstly observe that

$$|\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}| \leq \sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| \sup_t \|\boldsymbol{\gamma}(t)\| = \tilde{o}_p(\delta_n^{*-1})\tilde{O}_p(\delta_n^*) = \tilde{o}_p(1).$$

Thus we have

$$n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}\frac{[\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}]^2}{1+\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}} \sim n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}[\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}]^2.$$

Let $\boldsymbol{\gamma}^{\mathsf{T}}(t) = (\gamma_1(t), \gamma_2(t), \cdots, \gamma_p(t))$ and $g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\} = (g_{i1}\{\boldsymbol{\beta}(t)\}, \cdots, g_{ip}\{\boldsymbol{\beta}(t)\}), i = 1, 2, \cdots, n.$ Then u-th component of $n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}[\boldsymbol{\gamma}^{\mathsf{T}}(t)g_i\{\boldsymbol{\beta}(t)\}]^2$ is

$$n^{-1}\sum_{i=1}^{n}\sum_{j,k=1}^{p} \gamma_j(t)\gamma_k(t)g_{iu}\{\boldsymbol{\beta}(t)\}g_{ij}\{\boldsymbol{\beta}(t)\}g_{ik}\{\boldsymbol{\beta}(t)\}$$

whose absolute value can be bounded by

$$\begin{aligned}
&|n^{-1}\sum_{i=1}^{n}\sum_{j,k=1}^{p} \gamma_j(t)\gamma_k(t)g_{iu}\{\boldsymbol{\beta}(t)\}g_{ij}\{\boldsymbol{\beta}(t)\}g_{ik}\{\boldsymbol{\beta}(t)\}| \\
\leq&(\sup_t \|\boldsymbol{\gamma}(t)\|)^2 \sup_t \max_i |g_{iu}\{\boldsymbol{\beta}(t)\}| \Big| n^{-1}\sum_{i=1}^{n}\sum_{j,k=1}^{p} g_{ij}\{\boldsymbol{\beta}(t)\}g_{ik}\{\boldsymbol{\beta}(t)\}\Big| \\
\leq&(\sup_t \|\boldsymbol{\gamma}(t)\|)^2 \sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| \Big| n^{-1}\sum_{i=1}^{n}(\sum_{j=1}^{p} g_{ij}\{\boldsymbol{\beta}(t)\})^2\Big| \\
\leq&C(\sup_t \|\boldsymbol{\gamma}(t)\|)^2 \sup_t \max_i \|g_i\{\boldsymbol{\beta}(t)\}\| \sup_t \frac{p}{n}\sum_{i=1}^{n} \|g_i\{\boldsymbol{\beta}(t)\}\|^2 \\
=&\tilde{O}_p\{(\delta_n')^2\}\tilde{o}_p(\delta_n^{*-1})\tilde{O}_p(1) = \tilde{o}_p(\delta_n'),
\end{aligned}$$

This means that the third term in (2.6.39) is of order $\tilde{o}_p(\delta_n')$. It then follows that

$$\boldsymbol{\gamma}(t) = \left\{n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}\right\}^{-1}\left\{n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\}\right\} + \tilde{o}_p(\delta_n'). \tag{2.6.40}$$

Lemma 8. *Under assumptions (C1)-(C3) and (C4)(i), in the sphere*

$$\left\{\boldsymbol{\beta}(t) : \sup_{t\in[a,b]} \|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)\| \leq \delta_n^*\right\},$$

the equation system (2.3.15) almost surely has root in

$$U_{\delta_n^*} = \{(\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)) : \sup_t \|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t) + \boldsymbol{\gamma}(t) + \boldsymbol{\nu}(t)\| \leq \delta_n^*\}.$$

And any solution is indeed a solution to the minimization problem (3.2).

Proof. Since we have already proved in Lemma 7 that for every $\boldsymbol{\beta}(t) \in \{\boldsymbol{\beta}(t) : \sup_{t\in[a,b]} \|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)\| \leq \delta_n^*\}$, the equation $Q_{1n}(\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t)) = 0$ almost surely has root $\boldsymbol{\gamma}(t) = \boldsymbol{\gamma}(\boldsymbol{\beta}(t)) = \tilde{O}(\delta_n')$, we only have to prove the following:

(a) For every $\boldsymbol{\beta}(t) \in \{\boldsymbol{\beta}(t) : \sup_t \|\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)\| \leq \delta_n^*\}$, $\boldsymbol{\nu}(t) = \boldsymbol{\nu}\{\boldsymbol{\beta}(t)\} = \tilde{O}(\delta_n^*)$ could be solved from the equations $Q_{2n}(\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)) = 0$.

(b) And there almost surely exists a solution $\tilde{\boldsymbol{\beta}}(t) \in U_{\delta_n^*}$ to the equation system (3.3).

(c) Any solution is indeed a solution to the minimization problem (3.2).

In order to prove (a), recall the expression in (2.6.40) and the asymptotic variance $\mathbf{B}(t)$ in Lemma 3, by the uniformly strong law of large numbers (SLLN), we have

$$C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\} g_i^{\intercal}\{\boldsymbol{\beta}(t)\} = \mathbf{B}(t) + \tilde{o}_p(1).$$

Thus

$$\begin{aligned}\boldsymbol{\gamma}\{\boldsymbol{\beta}(t)\} &= \left\{ n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\} g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\} \right\}^{-1} \left\{ n^{-1}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\} \right\} + \tilde{o}_p(\delta_n') \\ &= \mathbf{B}^{-1}(t) \left\{ \frac{1}{C_{n,\alpha_0,\eta}^2} \sum_{i=1}^{n} g_i\{\boldsymbol{\beta}(t)\} \right\} + \tilde{o}_p(\delta_n'),\end{aligned} \tag{2.6.41}$$

and

$$\boldsymbol{\gamma}\{\boldsymbol{\beta}_0(t)\} = \tilde{O}_p(n\delta_n / C_{n,\alpha_0,\eta}^2) = \tilde{o}_p(n\delta_n^* / C_{n,\alpha_0,\eta}^2) = \tilde{o}_p(\delta_n'). \tag{2.6.42}$$

We have

$$\frac{\partial \boldsymbol{\gamma}\{\boldsymbol{\beta}(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} = \mathbf{B}^{-1}(t) \left\{ C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} \frac{\partial g_i\{\boldsymbol{\beta}(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} \right\} + \tilde{o}_p(\delta_n'). \tag{2.6.43}$$

Because the uniformly SLLN gives $n^{-1}\sum_{i=1}^{n} \frac{\partial g_i\{\boldsymbol{\beta}_0(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} = \mathbf{A}(t) + \tilde{o}_p(1)$ where $\mathbf{A}(t) = \boldsymbol{\Gamma}(t) f(t)$, we have the following

$$\begin{aligned}\frac{\partial \boldsymbol{\gamma}\{\boldsymbol{\beta}_0(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} &= \mathbf{B}^{-1}(t) \left\{ C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} \frac{\partial g_i\{\boldsymbol{\beta}_0(t)\}}{\partial \boldsymbol{\beta}^{\mathsf{T}}(t)} \right\} + \tilde{o}_p(\delta_n') \\ &= n C_{n,\alpha_0,\eta}^{-2} \mathbf{B}^{-1}(t) \mathbf{A}(t) + \tilde{o}_p(\delta_n').\end{aligned} \tag{2.6.44}$$

Let $S\{\boldsymbol{\beta}(t)\} = n^{-1}\sum_{i=1}^{n} \frac{\partial g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}/\partial \boldsymbol{\beta}(t)}{1+\boldsymbol{\gamma}^{\mathsf{T}}\{\boldsymbol{\beta}(t)\} g_i\{\boldsymbol{\beta}(t)\}}$, then

$$Q_{2n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)\} = S\{\boldsymbol{\beta}(t)\}\boldsymbol{\gamma}\{\boldsymbol{\beta}(t)\} + C^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t). \tag{2.6.45}$$

For the taylor expansion of $Q_{2n}\{\boldsymbol{\beta}(t), \boldsymbol{\gamma}(t), \boldsymbol{\nu}(t)\}$ at $\boldsymbol{\beta}_0(t)$, we need the following:

$$\begin{aligned} S\{\boldsymbol{\beta}(t)\} &= n^{-1}\sum_{i=1}^{n} \frac{\partial g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}/\partial\boldsymbol{\beta}(t)}{1+\boldsymbol{\gamma}^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}g_i\{\boldsymbol{\beta}(t)\}} \\ &= n^{-1}\sum_{i=1}^{n} \frac{\partial g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}}{\partial\boldsymbol{\beta}(t)} \left\{1 - \frac{\boldsymbol{\gamma}^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}g_i\{\boldsymbol{\beta}(t)\}}{1+\boldsymbol{\gamma}^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}g_i\{\boldsymbol{\beta}(t)\}}\right\} \\ &= n^{-1}\sum_{i=1}^{n} \frac{\partial g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}}{\partial\boldsymbol{\beta}(t)} + \tilde{O}_p(\delta_n'), \end{aligned} \tag{2.6.46}$$

which implies that

$$S\{\boldsymbol{\beta}_0(t)\} = \mathbf{A}(t) + \tilde{O}_p(\delta_n'). \tag{2.6.47}$$

Hence we have

$$\frac{\partial S\{\boldsymbol{\beta}(t)\}}{\partial\boldsymbol{\beta}^{\mathsf{T}}(t)} = n^{-1}\sum_{i=1}^{n} \frac{\partial^2 g_i^{\mathsf{T}}\{\boldsymbol{\beta}(t)\}}{\partial\boldsymbol{\beta}^{\mathsf{T}}(t)\partial\boldsymbol{\beta}(t)} + \tilde{O}_p(\delta_n'), \tag{2.6.48}$$

$$\frac{\partial S\{\boldsymbol{\beta}_0(t)\}}{\partial\boldsymbol{\beta}^{\mathsf{T}}(t)} = \mathbb{E}\frac{\partial^2 g_i^{\mathsf{T}}\{\boldsymbol{\beta}_0(t)\}}{\partial\boldsymbol{\beta}^{\mathsf{T}}(t)\partial\boldsymbol{\beta}(t)} + O_p(\delta_n') := \mathbf{D}(t) + \tilde{O}_p(\delta_n'). \tag{2.6.49}$$

Let $W\{\boldsymbol{\beta}(t)\} = S\{\boldsymbol{\beta}(t)\}\boldsymbol{\gamma}\{\boldsymbol{\beta}(t)\}$ and define $S\{\boldsymbol{\beta}(t)\} = (\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_p)$ where $\mathbf{S}_j$ is the j-th column of $S\{\boldsymbol{\beta}(t)\}$. Then by (2.6.42), (2.6.44), (2.6.47), (2.6.49) and the assumption about

$\boldsymbol{\beta}(t)$ we have

$$\begin{aligned} W\{\boldsymbol{\beta}(t)\} &= W\{\boldsymbol{\beta}_0(t)\} + S\{\boldsymbol{\beta}_0(t)\}\frac{\partial \boldsymbol{\gamma}\{\boldsymbol{\beta}_0(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)}(\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)) \\ &\quad + \sum_{j=1}^{p} \frac{\partial \mathbf{S}_j}{\partial \boldsymbol{\beta}^{\intercal}(t)}\boldsymbol{\gamma}_j\{\boldsymbol{\beta}_0(t)\}(\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)) + \tilde{O}_p\{(\delta_n^*)^2\} \\ &= \{\mathbf{A}(t) + \tilde{O}_p(\delta_n')\}\tilde{o}_p(\delta_n') \\ &\quad + \{[\mathbf{A}(t) + \tilde{O}_p(\delta_n')][nC_{n,\alpha_0,\eta}^{-2}\mathbf{B}^{-1}(t)\mathbf{A}(t) + \tilde{o}_p(\delta_n')] \\ &\quad + [\mathbf{D}(t) + \tilde{O}_p(\delta_n')]\tilde{o}_p(\delta_n')\}[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + \tilde{O}_p\{(\delta_n^*)^2\} \\ &= nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + \tilde{o}_p(\delta_n'). \end{aligned}$$

By plugging the above into (2.6.45), we get

$$0 = nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + C^{\intercal}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t) + \tilde{o}_p(\delta_n'). \tag{2.6.50}$$

Since $\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)$ is invertible, by multiplying $\mathbf{C}(t)\{\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)\}^{-1}$ on both side of (2.6.50) we have

$$0 = nC_{n,\alpha_0,\eta}^{-2}\mathbf{C}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + \mathbf{C}(t)\{\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)\}^{-1}C^{\intercal}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t) + \tilde{o}_p(\delta_n'). \tag{2.6.51}$$

From the third equation of the equation system (3.3),

$$\begin{aligned} 0 = H\{\boldsymbol{\beta}(t)\} &= H\{\boldsymbol{\beta}_0(t)\} + \mathbf{C}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + \tilde{o}(\delta_n') \\ &= \mathbf{C}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + \tilde{o}(\delta_n'), \end{aligned}$$

we have

$$\mathbf{C}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] = \tilde{o}(\delta_n'). \tag{2.6.52}$$

Combine (2.6.51) and (2.6.52),

$$\mathbf{C}(t)\{\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)\}^{-1}C^{\intercal}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t) = -nC_{n,\alpha_0,\eta}^{-2}\mathbf{C}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + o_p(\delta_n') = o_p(\delta_n'),$$

That is

$$\boldsymbol{\nu}(t) = \left\{\mathbf{C}(t)\{\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)\}^{-1}C^{\intercal}\{\boldsymbol{\beta}(t)\}\right\}^{-1} o_p(\delta_n') = o_p(\delta_n'). \tag{2.6.53}$$

Hence we proved (a).

For proving (b), from (2.6.50) and (2.6.53), we have

$$\begin{aligned} 0 &= nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + C^{\intercal}\{\boldsymbol{\beta}(t)\}\boldsymbol{\nu}(t) + o_p(\delta_n') \\ &= nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + o_p(\delta_n'), \end{aligned}$$

which implies that

$$0 = -\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + o_p(\delta_n^*). \tag{2.6.54}$$

Now consider the above equation (2.6.54) and define a function ϕ on the the unit disk in $\mathbb{R}^p$ by

$$\phi\left(\frac{\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)}{\delta_n^*}\right) = -\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0(t)] + o_p(\delta_n^*).$$

We know that ϕ is a continuous function on the unit disk. Also we have

$$\begin{aligned}
&\delta_n^{*-1}[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]^\intercal\phi\left(\frac{\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)}{\delta_n^*}\right)\\
=&-\delta_n^{*-1}[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]^\intercal\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]+o_p(\delta_n^*).
\end{aligned}$$

Hence on the circle $\|\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)\|=\delta_n^*$, we have

$$\begin{aligned}
&\delta_n^{*-1}[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]^\intercal\phi\left(\frac{\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)}{\delta_n^*}\right)\\
=&-\delta_n^{*-1}[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]^\intercal\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t)-\boldsymbol{\beta}_0(t)]+o_p(\delta_n^*)\\
\leq&-\delta_n^*\tau_0(t)+o_p(\delta_n^*)<0,\ \text{if } n \text{ big enough},
\end{aligned}$$

where $\tau_0(t)>0$ is the smallest eigenvalue of $\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)$, which is positive definite. Thus by the lemma in [AS58], there exists a point $\tilde{\boldsymbol{\beta}}(t)\in U_{\delta_n^*}$ and $\phi\{\frac{\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}_0(t)}{\delta_n^*}\}=0$, which means $\tilde{\boldsymbol{\beta}}(t)$ is a solution to the equation system (3.3).

Next we have to prove (c). Assuming that $\tilde{\boldsymbol{\beta}}(t)$ is a solution in $U_{\delta_n^*}$, we let $\boldsymbol{\beta}(t)$ be a point in a neighborhood of $\tilde{\boldsymbol{\beta}}(t)$ contained in $U_{\delta_n^*}$ such that $H\{\boldsymbol{\beta}(t)\}=0$ and $\|\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}(t)\|>\delta>0$. Then by expanding $l_0\{\boldsymbol{\beta}(t)\}$ at $\tilde{\boldsymbol{\beta}}(t)$ we have

$$\begin{aligned}
l_0\{\boldsymbol{\beta}(t)\}-l_0\{\tilde{\boldsymbol{\beta}}(t)\}=&\frac{\partial l_0\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial\boldsymbol{\beta}^\intercal(t)}[\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}(t)]\\
&+\frac{1}{2}[\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}(t)]^\intercal\frac{\partial^2 l_0\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)\partial\boldsymbol{\beta}^\intercal(t)}[\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}(t)],
\end{aligned}\tag{2.6.55}$$

where $\beta^*(t)\in U_{\delta_n^*}$. We wish to show that

$$l_0\{\boldsymbol{\beta}(t)\}-l_0\{\tilde{\boldsymbol{\beta}}(t)\}>0.$$

Next, we approximate the two terms on the right side of (2.6.55): For the first term, note that

$$\begin{aligned}\frac{\partial l_0\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)} &= \sum_{i=1}^{n} \frac{1}{1+\boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}g_i\{\tilde{\boldsymbol{\beta}}(t)\}} g_i^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}\frac{\partial \boldsymbol{\gamma}\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)} \\ &+ \sum_{i=1}^{n} \frac{1}{1+\boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}g_i\{\tilde{\boldsymbol{\beta}}(t)\}} \boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}\frac{\partial g_i\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)} \\ &= \sum_{i=1}^{n} \frac{1}{1+\boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}g_i\{\tilde{\boldsymbol{\beta}}(t)\}} \boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}\frac{\partial g_i\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)} \\ &= n\boldsymbol{\gamma}^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}S^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\} = nW^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}.\end{aligned} \tag{2.6.56}$$

By (2.6.45), we have

$$W^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\} = -\tilde{\boldsymbol{\nu}}^{\intercal}(t)C\{\tilde{\boldsymbol{\beta}}(t)\}. \tag{2.6.57}$$

From the taylor expansion of $H\{\boldsymbol{\beta}(t)\}$ at $\tilde{\boldsymbol{\beta}}(t)$, we have

$$0 = H\{\boldsymbol{\beta}(t)\} - H\{\tilde{\boldsymbol{\beta}}(t)\} = C\{\tilde{\boldsymbol{\beta}}(t)\}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] + \tilde{o}(\delta_n^*),$$

from which we could obtain

$$C\{\tilde{\boldsymbol{\beta}}(t)\}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] = \tilde{o}(\delta_n^*). \tag{2.6.58}$$

Thus, for the first term of (2.6.55), combining (2.6.56)-(2.6.58) we have

$$\begin{aligned}&\frac{\partial l_0\{\tilde{\boldsymbol{\beta}}(t)\}}{\partial \boldsymbol{\beta}^{\intercal}(t)}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] = nW^{\intercal}\{\tilde{\boldsymbol{\beta}}(t)\}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] \\ &= -n\tilde{\boldsymbol{\nu}}^{\intercal}(t)C\{\tilde{\boldsymbol{\beta}}(t)\}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] = -n^2 C_{n,\alpha_0,\eta}^{-2}\tilde{o}_p\{(\delta_n^*)^2\}.\end{aligned} \tag{2.6.59}$$

For the second term of (2.6.55), we have

$$\begin{aligned}\frac{\partial^2 l_0\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)\partial\boldsymbol{\beta}^\intercal(t)} &= n\frac{\partial W^\intercal\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)} = n\left\{\frac{\partial\boldsymbol{\gamma}^\intercal\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)}S^\intercal\{\beta^*(t)\} + \boldsymbol{\gamma}\{\beta^*(t)\}\frac{\partial S^\intercal\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)}\right\}\\ &= n[nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t) + \tilde{o}_p(\delta_n')][\mathbf{A}(t) + \tilde{O}_p(\delta_n')]\\ &\quad + n\tilde{O}_p(\delta_n')[\mathbf{D}(t) + \tilde{O}_p(\delta_n')]\\ &= n\{nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t) + \tilde{O}_p(\delta_n')\}.\end{aligned}$$

It follows that

$$\begin{aligned}&\frac{1}{2}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)]^\intercal\frac{\partial^2 l_0\{\beta^*(t)\}}{\partial\boldsymbol{\beta}(t)\partial\boldsymbol{\beta}^\intercal(t)}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)]\\ =&\frac{1}{2}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)]^\intercal n\{nC_{n,\alpha_0,\eta}^{-2}\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t) + \tilde{O}_p(\delta_n')\}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] \qquad (2.6.60)\\ =&\frac{n^2}{2C_{n,\alpha_0,\eta}^2}[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)]^\intercal\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)[\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)] + \frac{n^2}{C_{n,\alpha_0,\eta}^2}\tilde{o}_p\{(\delta_n^*)^2\}.\end{aligned}$$

Hence, plugging (2.6.59) and (2.6.60) into (2.6.55), we have

$$\begin{aligned}&l_0\{\boldsymbol{\beta}(t)\} - l_0\{\tilde{\boldsymbol{\beta}}(t)\}\\ &= \frac{n^2}{C_{n,\alpha_0,\eta}^2}\left\{\frac{1}{2}(\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t))^\intercal\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)(\boldsymbol{\beta}(t) - \tilde{\boldsymbol{\beta}}(t)) + \tilde{o}_p\{(\delta_n^*)^2\}\right\}\\ &\geq \frac{n^2}{C_{n,\alpha_0,\eta}^2}(\delta_n^*)^2\{\frac{1}{2}\tau_0(t) + \tilde{o}_p(1)\} > 0, \text{ if } n \text{ big enough},\end{aligned}$$

where $\tau_0(t) > 0$ is the smallest eigenvalue of $\mathbf{A}(t)\mathbf{B}^{-1}(t)\mathbf{A}(t)$, which is positive definite. □

Chapter 3

Unified simultaneous empirical likelihood ratio tests for functional linear models and the phase transition from sparse to dense functional data

3.1 Introduction

In this chapter, we continue to consider the same model (2.1.1) as we discussed in Chapter 2. And we are interested in the same hypothesis testing problem as in (2.1.2),

$$H_0 : H\{\boldsymbol{\beta}_0(\cdot)\} = 0 \quad \text{vs} \quad H_1 : H\{\boldsymbol{\beta}_0(\cdot)\} \neq 0. \tag{3.1.1}$$

But instead of testing the coefficient functions at a fixed point t as in Chapter 2, we would like to test the functions simultaneously on the whole support $[a, b]$.

In this chapter, we propose nonparametric test based on the pointwise empirical likelihood ratio test in Chapter 2, to test (2.1.2) simultaneously. Since in Chapter 2, we showed the EL-based pointwise tests enjoy a nice self-normalizing property such that both sparse and dense functional data can be treated under a unified framework, the simultaneous testing

procedure to be developed here can also treat all types of functional data with different denseness in a unified way.

To investigate the power of the tests, we consider the same local alternatives (2.1.3) as in Chapter 2 for the entire functions $\boldsymbol{\beta}_0(\cdot)$ simultaneously

$$H_{1n} : H\{\boldsymbol{\beta}_0(\cdot)\} = b_n \mathbf{d}(\cdot), \tag{3.1.2}$$

For the sparse data with $\eta = 0$, it is also known that the EL method using a global bandwidth h [CZ10] can detect alternatives of order $b_n = n^{-1/2}h^{-1/4}$ for simultaneous test, which is also larger than $n^{-1/2}$. Similarly as in the pointwise case in Chapter 2, for dense data with $\eta > 0$, the detectable order b_n is still largely unknown. This leads to the same key interest in this chapter as in the last chapter, understanding the effect of η on b_n. We use the same principle to get the optimal b_n by maximizing the power of the test (i.e., minimizing the order of b_n) while controlling the type I error at the desired level. Under some mild conditions, we find that, for the simultaneous test, b_n is larger than $n^{-1/2}$ for $\eta \leq 1/16$ and equals to $n^{-1/2}$ for $\eta > 1/16$. The transition points $1/16$ will be still refereed as η_0 as in the pointwise case for this simultaneous test. Once $\eta > \eta_0$, with a properly chosen bandwidth, the proposed tests can detect a signal at a parametric rate. This phase transition result echoes the similar phenomena discovered by [LH10] for estimation problems.

The rest of the chapter is organized as follows. We propose the unified simultaneous test in Section 3.2 where we investigate the asymptotic distributions of the test statistic under both the null and local alternatives, and the transition phases for b_n. Simulation studies are presented in Section 3.3, followed by two real data analysis examples, one for sparse and one for dense functional data, in Section 3.4. All the technical details are relegated to the

Section 3.5.

3.2 A unified simultaneous test

We assume the same regularity conditions (C1)-(C4) for kernel function, moments of the underlying processes, smoothness of the related functions and the selection of bandwidth as in 2.2.2 in Chapter 2.

We now consider a simultaneous test on H_0 in (3.1.1) for all $t \in [a, b]$. By Lemma 5 in Section 2.6.2 in Chapter 2

$$2\ell(t) = n^2 C_{n,\alpha_0,\eta}^{-2} H^\intercal\{\check{\boldsymbol{\beta}}(t)\}\mathbf{R}(t)H\{\check{\boldsymbol{\beta}}(t)\} + \tilde{o}_p(1).$$

Intuitively, $2\ell(t)$ measures the distance between $H\{\boldsymbol{\beta}_0(t)\}$ and 0 at any $t \in [a, b]$. To test the hypothesis (3.1.1) simultaneously, we propose a Cramér-von Mises type test statistic

$$T_n = \int_a^b 2\ell(t)w(t)dt, \tag{3.2.3}$$

where $w(\cdot)$ is a known probability density function. The construction of T_n allows us to borrow information across the time domain and yield a more powerful test than the pointwise test. Similar constructions were used by [HM93] and [CZ10]. The weight function $w(t)$ is a subjective choice of the practitioner. The most commonly used weight function is a uniform density to put equal weights on all points, but if there is prior knowledge on the importance of a particular subinterval one can change $w(t)$ to put more weights on the important subinterval.

3.2.1 Null distribution and local power

By the asymptotic decomposition of $2\ell(t)$ in Proposition 3 in Chapter 2, we need to first understand the covariance structure of the process $\mathbf{U}_n(t)$ in order to investigate the distribution of T_n.

Proposition 4. *Under Conditions (C1)-(C4) and* H_0, *Cov*$\{\mathbf{U}_n(s), \mathbf{U}_n(t)\} = \mathbf{\Sigma}_n(s,t) \times \{1 + o_p(1)\}$ *where*

$$\mathbf{\Sigma}_n(s,t) = \begin{cases} \mu_{20}^{-1} K^{(2)}(\frac{s-t}{h})\mathbf{I}_q, & \text{if } m^2h \to 0, \\ \mathbf{I}_q I(s=t) + mh\mathbf{\Sigma}_0(s,t)I(s \neq t) & \text{if } m^2h \to \infty \text{ and } mh \to 0, \\ \mathbf{\Sigma}_0(s,t), & \text{if } mh \to \infty, \end{cases}$$

$K^{(2)}(x) = \int K(y)K(x-y)dy$ *and* $\mathbf{\Sigma}_0(s,t) = \mathbf{G}(s)\mathbf{\Gamma}(s,t)\mathbf{G}^\intercal(t)\Omega(s,t)f(s)f(t)$.

Obviously, the leading term in the covariance of $\mathbf{U}_n(t)$ is different under different asymptotic scenarios. In the second case in the expression of $\mathbf{\Sigma}_n(s,t)$, the $\mathbf{I}_q I(s=t)$ term seems to dominate but is only non-zero in an area with Lebesgue measure 0; the $mh\mathbf{\Sigma}_0(s,t)I(s \neq t)$ term is nonzero almost everywhere and produces the leading order variance of T_n in this case.

Suppose the covariance function $\mathbf{\Sigma}_n(s,t)$ has the following spectral decomposition [Bal60]

$$\mathbf{\Sigma}_n(s,t) = \sum_{k=1}^{\infty} \gamma_{nk} \boldsymbol{\phi}_{nk}(s) \boldsymbol{\phi}_{nk}^\intercal(t) \text{ for any } s,t \in [a,b],$$

where $\gamma_{n1} \geq \gamma_{n2} \geq \cdots \geq 0$ are the ordered eigenvalues and $\boldsymbol{\phi}_{n1}(t), \phi_{n2}(t), \cdots$ are the associated eigenfunctions. The eigenfunctions are vector valued orthonormal functions sat-

isfying $\int_a^b \boldsymbol{\phi}_{nk}^{\intercal}(t)\boldsymbol{\phi}_{nl}(t)w(t)dt = \delta_k^l$ where $\delta_k^l = 1$ if $k = l$ and 0 otherwise. Even though the eigenvalues γ_{nk} change under different asymptotic scenarios, it is easy to verify that $\sum_{k=1}^{\infty}\gamma_{nk} = \mathrm{tr}\{\int \boldsymbol{\Sigma}_n(t,t)w(t)dt\} = q$ for all cases in Proposition 4. Also note that in the third case of Proposition 4, $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0$ does not depend on n and therefore $\gamma_{nk} \equiv \gamma_k$ and $\boldsymbol{\phi}_{nk}(t) \equiv \boldsymbol{\phi}_k(t)$ for all k.

To establish the asymptotic distribution of T_n, we need all the conditions in Chapter 2 with replacing the condition (C4)(ii) by

(C4)(ii'): $2(1+\eta)/17 < \alpha_0$ if $\eta \in [0, 1/8]$ and $1/8 < \alpha_0 < \eta$ if $\eta > 1/8$.

Under the null hypothesis, we can define a q-dimensional Gaussian process $\mathbf{U}(t)$, with mean $\mathbf{0}$ and covariance $\mathrm{Cov}(\mathbf{U}(s), \mathbf{U}(t)) = \boldsymbol{\Sigma}_n(s,t)$, as a counterpart of the process $\mathbf{U}_n(t)$. We will show that the limiting distribution of T_n is the same as that of $Z_n = \int_a^b \mathbf{U}^{\intercal}(t)\mathbf{U}(t)w(t)dt$, which follows a χ^2-mixture distribution. This result is described in the following theorem, the proof of which is provided in the Section 3.5.1.

Theorem 2. *Under H_0 in (3.1.1) and Conditions (C1)-(C3), (C4)(i) and (C4)(ii'), $T_n \stackrel{d}{=} Z_n \times \{1 + o_p(1)\}$, where $Z_n \stackrel{d}{=} \sum_{k=1}^{\infty}\gamma_{nk}\chi^2_{1,k}$ and $\chi^2_{1,k}$, $k = 1, 2, \ldots$, are independent chi-square random variables with one degree of freedom.*

Remark 5. *The asymptotic χ^2-mixture distribution in Theorem 2 is quite different from the asymptotic normal distribution for classic empirical likelihood ratio tests for independent data, time series or sparse longitudinal data [CHL03, CZ10]. In fact, for dense functional data, our calculation shows the $\mathbb{E}\{(T_n - \mathbb{E}T_n)^4\} \neq 3var^2(T_n)$, and hence T_n can behave quite differently from a Gaussian variable. However, for sparse or moderately dense functional data with $\eta \leq 1/16$, the χ^2-mixture is also asymptotically normal. This result is collected in the following corollary, the proof of which is given in Section 3.5.1.*

Corollary 1. *Under the same conditions as those in Theorem 2, if* $\eta \leq 1/16$*, we have*

$$h^{-1/2}(T_n - q) \xrightarrow{d} N(0, q\sigma_0^2)$$

where $\sigma_0^2 = 2\mu_{20}^{-2} \int_a^b w^2(t)dt \int_{-2}^2 \{K^{(2)}(u)\}^2 du$.

Corollary 1 makes a connection between our general results in Theorem 2 with the classic results. The null distribution of T_n is different under different asymptotic scenarios and may depend on some unknown quantities such as γ_{nk}, which makes it difficult to use in practice. In the next subsection, we will propose a bootstrap method unanimously applicable to all types of functional data to estimate this null distribution. Next, we study the power of the simultaneous test under the local alternatives.

Theorem 3. *Suppose that the local alternative hypothesis in (3.1.2) holds and Conditions (C1)-(C3), (C4)(i) and (C4)(ii') are satisfied.*

(a) If $\eta \leq 1/16$ *and* $b_n = n^{-1/2}(m^2h)^{-1/4}$*, then* $h^{-1/2}(T_n - q) \xrightarrow{d} N(\mu_0, q\sigma_0^2)$*, where* $\mu_0 = \int_a^b \mathbf{d}^\intercal(t)\mathbf{R}(t)\mathbf{d}(t)w(t)dt$ *and* σ_0^2 *is defined in Corollary 1.*

(b) If $1/16 < \eta \leq 1/8$*,* $\alpha_0 < 2\eta$ *and* $b_n = n^{-1/2+\epsilon}$ *for an arbitrarily small* $\epsilon > 0$*, then* $\sigma_1^{-1}(T_n - q - nb_n^2 mh\mu_0) \xrightarrow{d} N(0,1)$ *where* $\sigma_1^2 = 4nb_n^2(mh)^2\mu_1$ *and*

$$\mu_1 = \int_a^b \int_a^b \mathbf{d}^\intercal(t)\mathbf{R}^{1/2}(t)\mathbf{\Sigma}_0(t,s)\mathbf{R}^{1/2}(s)\mathbf{d}(s)w(t)w(s)dtds.$$

(c) If $\eta > 1/8$ *and* $b_n = n^{-1/2}$*, let* $u_k = \int_a^b [\mathbf{R}^{1/2}(t)\mathbf{d}(t)]^\intercal \boldsymbol{\phi}_k(t)w(t)dt$*. Then* $T_n \xrightarrow{d} \sum_{k=1}^\infty \gamma_k \chi_{1,k}^2\Big(u_k^2/\gamma_k\Big)$.

We can use Theorem 3 to examine the power and size of detectable signals of the simultaneous test under different scenarios. We use the same principle (2.3.18) in Chapter 2 to determine the optimal rate for b_n. When $\eta \leq 1/16$, following part (a) in Theorem 3, the asymptotic power of the test is $\mathscr{B}(\mathbf{d}) = \Phi\Big(-z_\alpha + \mu_0/\sqrt{q}\sigma_0\Big)$ where μ_0 and σ_0 are defined in Theorem 3 and $\Phi(\cdot)$ is the CDF of a standard normal distribution. The test has nontrivial powers for signals of size $b_n = n^{-1/2}(m^2h)^{-1/4}$. Under the constraints (C4)(i) and (C4)(ii') on h, b_n attains its minimum at $h_* = n^{-2(1+\eta+\delta)/17}$ for any arbitrary small $\delta > 0$ such that $b_n = n^{-8(1+\eta)/17+\delta/34}$. By letting $\delta \to 0$, the optimal detectable order is $b_n^* = n^{-8(1+\eta)/17}$.

When $1/16 < \eta \leq 1/8$, by our calculations in Proposition 4 and Theorem 2 the null distribution of T_n is a χ^2 mixture with mean $(\sum_{k=1}^{\infty}\gamma_{nk})\times\{1+o(1)\} = q\times\{1+o(1)\}$ and variance $(2\sum_k \gamma_{nk}^2)\times\{1+o(1)\} = \mathrm{tr}\{\iint \mathbf{\Sigma}_n^2(s,t)\, w(s)w(t)dsdt\}\times\{1+o(1)\} = O(mh)$. Therefore, the threshold for an α-level test is of the form $q + c_{n,\alpha}$, where $c_{n,\alpha} \leq (2\sum_k \gamma_{nk}^2/\alpha)^{1/2} = O(mh)$ by Chebyshev's inequality. By part (b) of Theorem 3, the asymptotic power is

$$\mathscr{B}(\mathbf{d}) = \Phi\Big(-\frac{c_{n,\alpha}}{2\sqrt{n}b_n mh\sqrt{\mu_1}} + \frac{\mu_0}{2\sqrt{\mu_1}}\sqrt{n}b_n\Big) \to 1,$$

for $b_n = n^{-1/2+\epsilon}$ with an arbitrarily small $\epsilon > 0$. This also means that the test has nontrivial powers for signals of size $b_n^* = n^{-1/2}$.

Similarly, the power of the test under case (c) is

$$\mathscr{B}(\mathbf{d}) = P\Big(\sum_{k=1}^{\infty}\gamma_k\chi_{1,k}^2\big(u_k^2/\gamma_k\big) > q + c_\alpha\Big)$$

where $q + c_\alpha$ is the α-th quantile of $\sum_{k=1}^{\infty}\gamma_k\chi_{1,k}^2$. In this case, $\mathscr{B}(\mathbf{d})$ is a constant as long as $\mathbf{d}(t)$ is a fixed non-zero function, which implies that the test has a non-trivial power if

$b_n = n^{-1/2}$. Combining parts (b) and (c), the optimal detectable order of the simultaneous test is $b_n^* = n^{-1/2}$ when $\eta > 1/16$.

Note that the optimal detectable order for the simultaneous test is smaller than that of the pointwise test we obtained in Chapter 2 when $\eta \leq 1/8$. This is understandable because the simultaneous test borrow information over the entire time domain and is more powerful. Both the pointwise and simultaneous tests can detect signals of root-n order for dense functional data with $\eta > 1/8$.

3.2.2 Wild bootstrap procedure

The asymptotic distributions of T_n are different for sparse and dense functional data, but the boundary between different scenarios is defined only in the asymptotic sense, making different asymptotic scenarios very difficult to distinguish in practice. To unify the inference procedure, we propose a wild bootstrap procedure [Mam93]. Some residual based bootstrap procedures have also been proposed in [Far97] and [ZC07] for dense functional data, but the consistency of such procedures was not investigated. The proposed bootstrap procedure consists of the following steps:

Step 1: Generating bootstrap samples $\{Y_{ij}^{*(b)}, t_{ij}^{(b)}, \mathbf{X}_{ij}^{(b)}\}_{b=1}^{B}$ according to the following model:

$$Y_{ij}^* = \tilde{\boldsymbol{\beta}}^{\intercal}(t_{ij})\mathbf{X}_{ij} + \epsilon_{ij}^*.$$

where $\tilde{\boldsymbol{\beta}}(t_{ij})$ is the solution of the estimating equations in (2.3.15) in Chapter 2. The residual vector $\boldsymbol{\epsilon}_i^* = (\epsilon_{i1}^*, \cdots, \epsilon_{im_i}^*)^{\intercal}$ is generated from an m_i-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance $\hat{\boldsymbol{\Omega}}_i = (\hat{\Omega}(t_{ij}, t_{ik}))_{j,k=1}^{m_i}$ where $\hat{\Omega}(t,s)$ is a consistent estimator of $\Omega(t,s)$ described in Section 2.4.2 in Chapter 2.

Step 2: Based on the b-th bootstrapped sample, compute a bootstrapped version of T_n, denoted as $T_n^{*(b)}$.

Step 3: Repeat Steps 1 and 2 a large integer B times to obtain B bootstrap values $\{T_n^{*(b)}\}_{b=1}^B$ and then find the $100(1-\alpha)\%$ quantile of $\{T_n^{*(b)}\}_{b=1}^B$, denoted as $\hat{t}_\alpha$. Reject the null hypothesis if $T_n > \hat{t}_\alpha$.

The following theorem justifies the above Bootstrap procedure

Theorem 4. *Let $\mathcal{X}_n = \{(Y_{ij}, X_{ij}, t_{ij}),\ j = 1, \ldots, m_i,\ i = 1, \ldots, n\}$ denotes the original data and $\mathscr{L}(T_n)$ be the asymptotic distribution of T_n under the null hypothesis. Under the same conditions as Theorem 2 and suppose $\hat{\Omega}(s,t)$ is a consistent covariance estimator, the conditional distribution of T_n^* given $\mathcal{X}_n$, $\mathscr{L}(T_n^*|\mathcal{X}_n)$ converges to $\mathscr{L}(T_n)$ almost surely.*

3.3 Simulation studies

For the simulation studies for simultaneous inference, we consider the same setup as in the simulation studies for the pointwise inference in Section 2.5 in Chapter 2. We considered two scenarios A and B, corresponding to two hypotheses on $\boldsymbol{\beta}(t)$. In scenario A, we used $H\{(z_1, z_2)^\intercal\} = z_1 - z_2$ to test

$$H_{0A} : \beta_1(\cdot) = \beta_2(\cdot) \quad \text{vs} \quad H_{1A} : \beta_1(\cdot) \neq \beta_2(\cdot),$$

where we set $\beta_1(t) = \frac{1}{2}\sin t$ and $\beta_2(t) = (\frac{1}{2} + a)\sin t$ for $a = 0, 0.1, 0.2, 0.3$ and 0.4 in (2.5.19) in Chapter 2 to evaluate the empirical size (when $a = 0$) and powers (when $a > 0$). In

scenario B, we set $H\{(z_1, z_2)^\intercal\} = z_2$ to test

$$H_{0B} : \beta_2(\cdot) = 0 \quad \text{vs} \quad H_{1B} : \beta_2(\cdot) \neq 0,$$

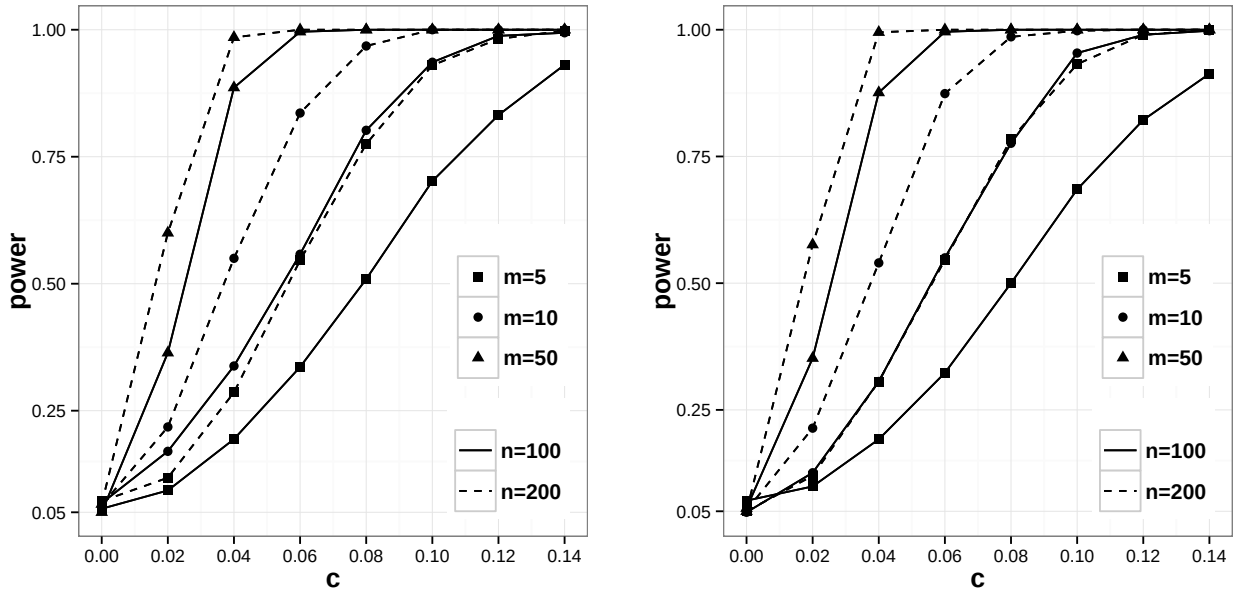
where we chose $\beta_1(t) = \frac{1}{2}\sin t$ and $\beta_2(t) = c$ for $c = 0, 0.02, 0.04, \cdots, 0.14$. In the construction of the test statistic T_n, we chose the weight function $w(t) = 1$ for $t \in (0, 1)$ and 0 otherwise. The covariance function was estimated by the quasi maximum likelihood method of [FHL07]. All simulation results below were based on 500 simulation replicates and the critical value of the test was estimated by 500 bootstrap samples in each simulation run. We performed the same bandwidth selection procedure in each bootstrap sample to take into account the extra variation in the test caused by bandwidth selection.

Table 3.1 summarizes the empirical sizes and powers for hypothesis H_{0A} at the 5% nominal level. It can be seen that the empirical sizes are reasonably controlled around the nominal level. As we expected, the empirical power increases as the increase of the sample size n and the number of repeated measurements m, which confirms our theoretical results in Section 3.2. In addition, the correlation ρ does not have a clear impact on the power, indicating that the proposed procedure is robust with respect to the covariance structure of the random error.

The simulation results for scenario B are illustrated in Figure 3.1. The results under $n = 100$ and $n = 200$ are represented by solid and dashed lines, respectively. We observed a very similar pattern as that under scenario A. The size is well controlled at the 5% nominal level and the power increases as the value of c increases. At each value of c, the power increases as we increase n or m.

Table 3.1: Empirical size and power for testing $H_{0A} : \beta_1(\cdot) = \beta_2(\cdot)$ under scenario A.

		$m = 5$		$m = 10$		$m = 50$	
a	n	ρ =0.2	ρ =0.5	ρ =0.2	ρ =0.5	ρ =0.2	ρ =0.5
0.0	100	0.062	0.058	0.064	0.048	0.070	0.054
	200	0.060	0.052	0.068	0.044	0.058	0.066
0.1	100	0.134	0.132	0.188	0.212	0.772	0.764
	200	0.224	0.228	0.388	0.344	0.984	0.966
0.2	100	0.344	0.406	0.676	0.708	1.000	1.000
	200	0.724	0.734	0.948	0.948	1.000	1.000
0.3	100	0.746	0.748	0.976	0.982	1.000	1.000
	200	0.974	0.974	0.998	1.000	1.000	1.000
0.4	100	0.962	0.960	1.000	1.000	1.000	1.000
	200	1.000	1.000	1.000	1.000	1.000	1.000



(a) $\rho = 0.2$ (b) $\rho = 0.5$

Figure 3.1: Empirical size and power for testing $H_{0B} : \beta_2(\cdot) = 0$ at the 5% nominal level under scenario B. The left panel is for $\rho = 0.2$ and the right panel is for $\rho = 0.5$.

3.4 Real data analysis

We applied our proposed methods to two real functional data sets, one is sparse and the other is dense.

3.4.1 CD4 data analysis

This data set was collected from a randomized double-blinded study of AIDS patients with advanced immune suppression (CD4 counts ≤ 50 cells/mm^3) conducted by the AIDS Clinical Trial Group (ACTG) Study 193A. Patients were randomly assigned to dual or triple combinations of HIV-1 reverse transcriptase inhibitors. Specifically, patients were randomized to one of four daily regimens containing 600mg of zidovudine: zidovudine alternating monthly with 400mg didanosine (treatment I); zidovudine plus 2.25mg of zalcitabine (treatment II); zidovudine plus 400mg of didanosine (treatment III); or zidovudine plus 400mg of didanosine plus 400mg of nevirapine (treatment IV). There was a total of 1309 patients included in the study and 325, 324, 330 and 330 patients were, respectively, assigned to treatments I-IV. Measurements of CD4 counts were collected at baseline and at 8-week intervals during follow-up. But due to various reasons, such as dropout and skipped visits, the repeated measurements were unbalanced. The number of repeated measurements during the first 40 weeks of follow-up varied from 1 to 9, with a median of 4. Thus, the data can be considered as sparse functional data. More details of the study can be found in [KAC$^+$98].

Our interest is to study the treatment effects on the CD4 counts. We consider the response variable to be log(CD4 counts + 1). To test for treatment effects, we set treatment IV as the baseline and defined three dummy variables T_1, T_2 and T_3 as indicators of treatments

I-III, respectively. Then, we fit the data with the following functional linear model:

$$\begin{aligned} Y_i(t_{ij}) =& \beta_0(t_{ij}) + \beta_1(t_{ij})T_{1i} + \beta_2(t_{ij})T_{2i} + \beta_3(t_{ij})T_{3i} \\ &+ \beta_4(t_{ij})\text{Age}_i(t_{ij}) + \beta_5(t_{ij})\text{Gender}_i + \beta_6(t_{ij})\text{PreCD4}_i + \epsilon_i(t_{ij}), \end{aligned}$$

for $i = 1, \cdots, 1309$ and $j = 1, \cdots, m_i$ where $Y(t) = \log(\text{CD4 counts} + 1)$ is the response, t is the time (in weeks). We also included Age, Gender and PreCD4 as the covariates in the model and allowed Age change over t.

To test for treatment effects, we first considered the global hypotheses

$$H_{01} : \beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0 \text{ vs } H_{11} : \text{at least one of } \beta_k(\cdot) \neq 0,\ k = 1, 2, 3.$$

We applied the proposed simultaneous test based on 1000 bootstrap replicates. The bandwidth was selected by the proposed procedure in Section 2.4. We got a p-value of < 0.001 indicating that the treatment effects are indeed significant. To further dissect differences between treatments, we conducted pairwise comparison among treatments. The results are summarized in Table 3.2. All the p-values for the pairwise comparisons except the one for comparing treatment II and III are less than 5%. The results indicate that pairwise differences in time effects between different treatment groups are statistically significant except for treatment II vs III.

3.4.2 Ergonomics data analysis

As part of a study of the body motions of automobile drivers, researchers at the Center for Ergonomics at the University of Michigan collected data on the motion of a single individual

Table 3.2: P-values for pairwise comparison among different treatment groups.

Comparison	Hypothesis	p-value
I vs II	$H_{02} : \beta_1(\cdot) = \beta_2(\cdot)$	0.040
I vs III	$H_{03} : \beta_1(\cdot) = \beta_3(\cdot)$	0.000
I vs IV	$H_{04} : \beta_1(\cdot) = 0$	0.000
II vs III	$H_{05} : \beta_2(\cdot) = \beta_3(\cdot)$	0.078
II vs IV	$H_{06} : \beta_2(\cdot) = 0$	0.000
III vs IV	$H_{07} : \beta_3(\cdot) = 0$	0.002

to 20 target locations within a test car. For each location, the researchers measured 3 times the angle formed at the right elbow between the upper and lower arms, which yielded a sample of size 20×3=60. The angle of each motion was recorded repeatedly from the start to the end of each test drive. The time period of each motion varied in length because of the targets being at different distances from the driver and the driver may reach them at different speeds. The objective of the study was to model the shape of the motion but not the speed at which it occurred. Thus in this study, t is used to represent the proportion, not the time, of the motion between the start and the end. See [Far97] and [SF04] for a more detailed description of this data set.

Let $Y(t)$ represent the angle at a proportion t for $t \in [0, 1]$. For a given motion, $Y(t)$ is observed on an equally spaced grid of points. Although the number of such points in the original data varies from observation to observation, the number of repeat measurements for each motion is 20 after imputation, which was considered as dense functional data as in [Zha11]. The purpose of our study was to find a model for predicting the right elbow angle curve $Y(t), t \in [0, 1]$ given the coordinates (c_x, c_y, c_z) of the target, where c_x represents the "left to right" direction, c_y represents the "close to far" direction, and c_z represents the "down to up" direction. The coordinates (c_x, c_y, c_z) of each of the 20 targets in the experiment were known and used as predictors in our model. [SF04] compared a linear model, a quadratic

model and a one-way ANOVA model. They found that a quadratic model of the following form fit the data adequately

$$\begin{aligned} Y_i(t_{ij}) =& \beta_1(t_{ij}) + c_{xi}\beta_2(t_{ij}) + c_{yi}\beta_3(t_{ij}) + c_{zi}\beta_4(t_{ij}) \\ &+ c_{xi}^2\beta_5(t_{ij}) + c_{yi}^2\beta_6(t_{ij}) + c_{zi}^2\beta_7(t_{ij}) \\ &+ c_{xi}c_{yi}\beta_8(t_{ij}) + c_{yi}c_{zi}\beta_9(t_{ij}) + c_{zi}c_{xi}\beta_{10}(t_{ij}) + \epsilon_i(t_{ij}). \end{aligned} \tag{3.4.4}$$

for $i = 1, \cdots, 60$ and $j = 1, \cdots, 20$.

We started with model (3.4.4), and tested each of the coefficient functions $\beta_k(t)$, $k = 1, \cdots, 10$ to check which term could be dropped from the model. Table 3.3 summarizes the p-values for testing each coefficient function. At the 5% significant level, we can see that $\beta_7(t), \beta_9(t)$ and $\beta_{10}(t)$ are not significant, suggesting to delete them from the quadratic model (3.4.4). We then obtained the final reduced model

$$\begin{aligned} Y(t) =& \beta_1(t) + c_x\beta_2(t) + c_y\beta_3(t) + c_z\beta_4(t) \\ &+ c_x^2\beta_5(t) + c_y^2\beta_6(t) + c_xc_y\beta_8(t) + \epsilon(t). \end{aligned}$$

From the above reduced model, we could see that the angle curve $Y(t)$ has a significant linear relationship with the "down to up" coordinate z, but a significant quadratic relationship with the "left to right" coordinate x and the "close to far" coordinate y. The model selected above is consistent with the model chosen by [Zha11].

Table 3.3: P-values for testing each coefficient function in the quadratic model (3.4.4).

Hypothesis	p-value	Hypothesis	p-value
$H_{01} : \beta_1(\cdot) = 0$	0.000	$H_{06} : \beta_6(\cdot) = 0$	0.032
$H_{02} : \beta_2(\cdot) = 0$	0.006	$H_{07} : \beta_7(\cdot) = 0$	0.050
$H_{03} : \beta_3(\cdot) = 0$	0.006	$H_{08} : \beta_8(\cdot) = 0$	0.004
$H_{04} : \beta_4(\cdot) = 0$	0.005	$H_{09} : \beta_9(\cdot) = 0$	0.080
$H_{05} : \beta_5(\cdot) = 0$	0.038	$H_{0,10} : \beta_{10}(\cdot) = 0$	0.109

3.5 Technical Details

This section contains the proofs for the main theorems in Section 3.2. Proofs for the propositions can be found in the next section.

3.5.1 Proofs of Main Theorems

3.5.1.1 Proof of Theorem 2

Proof of Theorem 2. We first prove the case with $\eta \in [0, 1/8]$, under which we choose the bandwidth $h = n^{-\alpha_0}$ from $2(1+\eta)/17 < \alpha_0 < 1 - \eta - 2/\lambda$. In this scenario, it is easy to see that $mh \to 0$. In this case, we have the decomposition for T_n, $T_n = T_{n1} + T_{n2}$, where

$$T_{n1} = C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} \int_a^b \xi_i^{\intercal}(t)\mathbf{G}^{\intercal}(t)\mathbf{G}(t)\xi_i(t)w(t)dt$$
$$T_{n2} = C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} \sum_{k\neq i}^{n} \int_a^b \xi_i^{\intercal}(t)\mathbf{G}^{\intercal}(t)\mathbf{G}(t)\xi_k(t)w(t)dt.$$

It then can be shown that

$$\begin{aligned}
\mathbb{E}(T_{n1}) &= q + qh\frac{m-\bar{r}}{\bar{r}\mu_{20}}\int_a^b f(t)w(t)dt + O(h^2) \\
\mathrm{Var}(T_{n1}) &= \Big\{q + qh\frac{m-\bar{r}}{\bar{r}\mu_{20}}\int_a^b f(t)w(t)dt\Big\}^2 + O(h^2 + 1/n) \\
&\quad - \Big\{q + qh\frac{m-\bar{r}}{\bar{r}\mu_{20}}\int_a^b f(t)w(t)dt\Big\}^2 + O(h^2) = O(h^2 + 1/n),
\end{aligned}$$

and $E(T_{n2}) = 0$,

$$\begin{aligned}
\mathrm{Var}(T_{n2}) &= 2qh\mu_{20}^{-2}\int_{-2}^{2}[K^{(2)}(u)]^2du\int_a^b w^2(t)dt \\
&\quad + 2(mh)^2\int_a^b\int_a^b \mathrm{tr}\{\mathbf{\Sigma}_0(t,s)\mathbf{\Sigma}_0(s,t)\}w(t)w(s)dtds + O(mh^2 + h/n).
\end{aligned}$$

Hence we have $\mathrm{Var}(T_{n1}) = O(h^2 + 1/n) = o\{\mathrm{Var}(T_{n2})\}$. It follows that

$$T_n - \mathbb{E}(T_n) = T_{n1} - \mathbb{E}(T_{n1}) + T_{n2} = T_{n2}\{1 + o_p(1)\}.$$

Thus, to study the asymptotic property of T_n, we only need to study that of T_{n2}.

In fact, we can write T_{n2} as

$$T_{n2} = \frac{1}{n}\sum_{i\neq k}^{n}\int_a^b \mathcal{Z}_i^{\intercal}(t)\mathcal{Z}_k(t)w(t)dt,$$

where $\mathcal{Z}_i(t) = \sqrt{mh}\mathbf{G}(t)\xi_i(t)$. Let $\mathcal{U}_n = \frac{1}{n-1}T_{n2} = \frac{2}{n(n-1)}\sum_{1\leq i<k\leq n}\mathcal{K}(\mathcal{Z}_i, \mathcal{Z}_k)$, where the symmetric kernel $\mathcal{K}(\mathcal{Z}_i, \mathcal{Z}_k) = \int_a^b \mathcal{Z}_i^{\intercal}(t)\mathcal{Z}_k(t)w(t)dt$. We define an operator $A_{\mathcal{K}}$ associated with the kernel $\mathcal{K}$ as $A_{\mathcal{K}}g(x) = \int_{-\infty}^{\infty}\mathcal{K}(x,y)g(y)dF(y)$, where F is the distribution of $\mathcal{Z}_i$. Then we have the associated eigenvalues and eigenfunctions, denoted as $\{\lambda_k, \psi_k\}_{k=1}^{\infty}$. By

U-statistic theory [Ser80], we have

$$n\mathcal{U}_n - \sum_{k=1}^{\infty} \lambda_k(\chi^2_{1,k} - 1) = o_p(1),$$

where $\{\chi^2_{1,k}\}_{k=1}^{\infty}$ are independent chi-square distributed random variables with 1 degree of freedom. That is $T_{n2} - \sum_{k=1}^{\infty} \lambda_k(\chi^2_{1,k} - 1) = o_p(1)$. Now we only need to prove that $\{\lambda_k\}_{k=1}^{\infty}$ is the same as $\{\gamma_{nk}\}_{k=1}^{\infty}$ from $\mathbf{\Sigma}$.

In fact, $\mathrm{Cov}(\mathcal{Z}_i(s), \mathcal{Z}_i(t)) = \mathbf{\Sigma}_n(s,t) = \sum_{k=1}^{\infty} \gamma_{nk}\boldsymbol{\phi}_{nk}(s)\boldsymbol{\phi}_{nk}^{\intercal}(t)$. Then we have the K-L representation of the random process $\mathcal{Z}(t) = \sum_{k=1}^{\infty} \xi_k^z \boldsymbol{\phi}_{nk}(t)$. Then

$$\begin{aligned}
A_{\mathcal{K}}\xi_m^x &= \int_{-\infty}^{\infty} \mathcal{K}(x,y)\xi_m^y dF(y) = \int_{-\infty}^{\infty}\int_a^b \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \xi_i^x \xi_j^y \boldsymbol{\phi}_{ni}^{\intercal}(t)\boldsymbol{\phi}_{nj}(t)w(t)\xi_m^y dt dF(y) \\
&= \int_a^b \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \xi_i^x \boldsymbol{\phi}_{ni}^{\intercal}(t)\boldsymbol{\phi}_{nj}(t)w(t)[\int_{-\infty}^{\infty} \xi_j^y \xi_m^y dF(y)]dt \\
&= \int_a^b \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \xi_i^x \boldsymbol{\phi}_{ni}^{\intercal}(t)\boldsymbol{\phi}_{nj}(t)w(t)\gamma_m \delta_{nj}^m dt \\
&= \gamma_{nm}\sum_{i=1}^{\infty} \xi_i^x \int_a^b \boldsymbol{\phi}_{ni}^{\intercal}(t)\boldsymbol{\phi}_{nm}(t)w(t)dt = \gamma_{nm}\sum_{i=1}^{\infty} \xi_i^x \delta_i^m = \gamma_{nm}\xi_m^x.
\end{aligned}$$

That is $\{\lambda_{nk}, \psi_{nk}\}_{k=1}^{\infty} = \{\gamma_k, \xi_k\}_{k=1}^{\infty}$. Thus we have $T_{n2} - \sum_{k=1}^{\infty} \gamma_{nk}(\chi^2_{1,k} - 1) = o_p(1)$, and then $T_n - \mathbb{E}(T_n) = T_{n2} + o_p(1) = \sum_{k=1}^{\infty} \gamma_{nk}(\chi^2_{1,k} - 1) + o_p(1)$. It follows that $T_n - q = \sum_{k=1}^{\infty} \gamma_{nk}(\chi^2_{1,k} - 1) + o_p(1)$. Since $\sum_{k=1}^{\infty} \gamma_{nk} = q$, we have $T_n = \sum_{k=1}^{\infty} \gamma_{nk}\chi^2_{1,k}\{1 + o_p(1)\}$.

We finally prove the result for the dense case, i.e. $\eta > 1/8$. In this case, we choose then bandwidth $h = n^{-\alpha_0}$ from $1/8 < \alpha_0 < \min\{\eta, 1/2 - 1/\lambda\}$. Under this scenario, we have $mh \to \infty$. By Lemma 9 in Section 3.5.2, we know $\mathbf{U}_n(t)$ asymptotically converges to a Gaussian process $\mathbf{U}(t;\eta)$ with mean 0 and covariance function $\mathbf{\Sigma}(s,t)$. Thus the limiting distribution of T_n is the same as the distribution of $Z = \int \mathbf{U}(t;\eta)^{\intercal}\mathbf{U}(t;\eta)w(t)dt$. We

only need to show the distribution of Z. To this end, using the following Karhunen-Loeve representation for $\mathbf{U}(t;\eta)$ [Bal60]

$$\mathbf{U}(t;\eta) = \sum_{k=1}^{\infty} \xi_k \boldsymbol{\phi}_k(t),$$

where $\xi_k = \int_a^b \mathbf{U}(t;\eta)^\intercal \boldsymbol{\phi}_k(t) w(t) dt$ are independent ($k = 1, 2, \cdots, \infty$) normal with mean 0 and variance γ_k. Here γ_k and $\boldsymbol{\phi}_k(t)$ are, respectively, the k-th ordered eigenvalue of $\boldsymbol{\Sigma}(s,t)$ and the corresponding eigenfunctions in $\mathbb{R}^q$. Then we have

$$Z = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \xi_k \xi_l \int_a^b \boldsymbol{\phi}_k(t)^\intercal \boldsymbol{\phi}_l(t) w(t) dt = \sum_{k=1}^{\infty} \xi_k^2.$$

Since ξ_k are independent $N(0, \gamma_k)$, we have $T_n \xrightarrow{d} Z = \sum_{k=1}^{\infty} \gamma_k \chi_{1,k}^2$. Thus by combining the above two cases together, we complete the proof of part (b). □

3.5.1.2 Proof of Corollary 1

Proof. From Theorem 2, T_n has the same distribution as $Z_n = \sum_{k=1}^{\infty} \gamma_{nk} \chi_{1,k}^2$. Thus we only need to show the asymptotical normality of $\sum_{k=1}^{\infty} \gamma_k \chi_{1,k}^2$. By Lyapunov central limit theorem, if the following condition hold

$$\sum_{k=1}^{\infty} \gamma_{nk}^4 / (\sum_{k=1}^{\infty} \gamma_{nk}^2)^2 \to 0, \tag{3.5.5}$$

Then we have

$$\frac{Z_n - \sum_{k=1}^{\infty} \gamma_{nk}}{\sqrt{2 \sum_{k=1}^{\infty} \gamma_{nk}^2}} \xrightarrow{d} N(0,1).$$

Using Proposition 4, $\mathbf{\Sigma}_n(s,t) = \mu_{20}^{-1} K^{(2)}(\frac{s-t}{h})\mathbf{I}_q$ and in particular $\mathbf{\Sigma}_n(t,t) = \mathbf{I}_q$, we find that $\sum_{k=1}^{\infty} \gamma_{nk} = \mathrm{tr}(\mathbf{\Sigma}) = q\int_a^b w(t)dt = q$ and

$$\sum_{k=1}^{\infty} \gamma_{nk}^2 = q\int_a^b \int_a^b \mu_{20}^{-2}\{K^{(2)}(\frac{s-t}{h})\}^2 w(s)w(t)dsdt = qh\sigma_0^2/2,$$

where σ_0^2 was defined in the corollary. Therefore, the conclusion in this Lemma holds. It remains to show the condition (3.5.5). Let $\gamma(s,t) = \mu_{20}^{-1}K^{(2)}(\frac{s-t}{h})$. Then

$$\begin{aligned}\sum_{k=1}^{\infty} \gamma_{nk}^4 &= q\int\int\int\int \gamma(s,t)\gamma(t,l)\gamma(l,m)\gamma(m,s)w(s)w(t)w(l)w(m)dsdtdldm \\ &= qh^3 C_0 \mu_{20}^{-4}\int_a^b w^4(t)dt\end{aligned}$$

where $C_0 = \int K^{(2)}(u_1)K^{(2)}(u_2)K^{(2)}(u_3)K^{(2)}(u_1+u_2+u_3)du_1du_2du_3$ is a constant. Thus the condition (3.5.5) holds. This completes the proof of this corollary. □

3.5.1.3 Proof of Theorem 3

Proof of Theorem 3. First notice that $2\ell(t) = n^2 C_{n,\alpha_0,\eta}^{-2}\tilde{\boldsymbol{\nu}}^{\intercal}(t)\mathbf{R}^{-1}(t)\tilde{\boldsymbol{\nu}}(t) + o_p(h^{1/2})$ and under local alternative,

$$\tilde{\boldsymbol{\nu}}(t) = -\mathbf{R}(t)\mathbf{C}(t)\mathbf{A}^{-1}(t)\big\{\frac{1}{n}\sum_{i=1}^{n} g_i\{\boldsymbol{\beta}_0(t)\}\big\} + \mathbf{R}(t)\mathbf{H}\{\boldsymbol{\beta}_0(t)\} + \tilde{o}_p(\delta_n).$$

We then define $\mathbf{U}_n^{+}(t) = \mathbf{G}(t)C_{n,\alpha_0,\eta}^{-1}\sum_{i=1}^{n}\xi_i(t) - nC_{n,\alpha_0,\eta}^{-1}\mathbf{R}^{1/2}(t)\mathbf{H}\{\boldsymbol{\beta}_0(t)\}$.

First considering the proof for part (a) with $0 \le \eta \le \eta_0 = 1/16$, under which we choose the bandwidth $h = n^{-\alpha_0}$ with $2(1+\eta)/17 < \alpha_0 < 1-\eta-2/\lambda$. In this scenario, we have

$m^2h \to 0$. We have

$$
\begin{aligned}
T_n = \int_a^b 2\ell(t)w(t)dt &= \int_a^b \mathbf{U}_n^{+\intercal}(t)\mathbf{U}_n^{+}(t)w(t)dt + o_p(h^{1/2}) \\
&= C_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n}\sum_{k=1}^{n} \int_a^b \xi_i^{\intercal}(t)\mathbf{G}^{\intercal}(t)\mathbf{G}(t)\xi_k(t)w(t)dt \\
&- 2nC_{n,\alpha_0,\eta}^{-2} \sum_{i=1}^{n} \int_a^b \xi_i^{\intercal}(t)\mathbf{G}^{\intercal}(t)\mathbf{R}^{1/2}(t)\mathbf{H}\{\boldsymbol{\beta}_0(t)\}w(t)dt \\
&+ n^2C_{n,\alpha_0,\eta}^{-2} \int_a^b \mathbf{H}^{\intercal}\{\boldsymbol{\beta}_0(t)\}\mathbf{R}(t)\mathbf{H}\{\boldsymbol{\beta}_0(t)\}w(t)dt + o_p(h^{1/2}) \\
&:= R_{n1} - 2R_{n2} + R_{n3} + o_p(h^{1/2}).
\end{aligned}
$$

Then by the result in Corollary 1, we have $h^{-1/2}\{R_{1n} - q\} \xrightarrow{d} N(0, q\sigma_0^2)$. And for R_{2n}, obviously we have $\mathbb{E}(R_{2n}) = 0$, and

$$
\begin{aligned}
\mathrm{Var}(R_{2n}) &= n^2b_n^2C_{n,\alpha_0,\eta}^{-4} \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j=1}^{m_i} \frac{1}{m_i^2}\mathbb{E}\int_a^b\int_a^b \mathbf{X}_{ij}^{\intercal}\mathbf{G}^{\intercal}(t)\mathbf{R}^{1/2}(t)\mathbf{d}(t) \\
&\quad \times \mathbf{X}_{il}^{\intercal}\mathbf{G}^{\intercal}(s)\mathbf{R}^{1/2}(s)\mathbf{d}(s)\epsilon_{ij}\epsilon_{il}K_h(t_{ij}-t)K_h(t_{il}-s)w(t)w(s)dtds \\
&= O(n^3b_n^2C_{n,\alpha_0,\eta}^{-4}) = O\{n(b_nmh)^2\}.
\end{aligned}
$$

Since in this case $b_n = (nm)^{-1/2}h^{-1/4}$, we have $\mathrm{Var}(R_{2n}) = O(mh^{3/2})$. Thus we have $h^{-1/2}R_{2n} \xrightarrow{p} 0$ since we have $mh^{1/2} \to 0$ under this case. And for R_{3n} which is non-random, we have $h^{-1/2}R_{3n} = \int_a^b \mathbf{d}^{\intercal}(t)\mathbf{R}(t)\mathbf{d}(t)w(t)dt$. Thus we have $h^{-1/2}(T_n - q) \xrightarrow{d} N(\mu_0, q\sigma_0^2)$, where $\mu_0 = \int_a^b \mathbf{d}^{\intercal}(t)\mathbf{R}(t)\mathbf{d}(t)w(t)dt$.

For part (b) with $1/16 < \eta \leq 1/8$, under which we choose the bandwidth $h = n^{-\alpha_0}$ with $2(1+\eta)/17 < \alpha_0 < 1 - \eta - 2/\lambda$. In this scenario, we also make to have $m^2h \to \infty$ and

$mh \to 0$. We write

$$\mathbf{U}_n^+(t) = C_{n,\alpha_0,\eta}^{-1} \sum_{i=1}^{n} \big\{ \mathbf{G}(t)\xi_i(t) - b_n \mathbf{R}^{1/2}(t)\mathbf{d}(t) \big\} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{Z}_i^+(t),$$

where $\mathcal{Z}_i^+(t) = \sqrt{mh}\big\{ \mathbf{G}(t)\xi_i(t) - b_n \mathbf{R}^{1/2}(t)\mathbf{d}(t) \big\}$. Then we have

$$\begin{aligned} T_n &= \int_a^b 2\ell(t)w(t)dt = \int_a^b \mathbf{U}_n^{+\intercal}(t)\mathbf{U}_n^+(t)w(t)dt + o_p(mh) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} \int_a^b \mathcal{Z}_i^{+\intercal}(t)\mathcal{Z}_k^+(t)w(t)dt + o_p(mh) := T_{n1}^+ + T_{n2}^+ + o_p(mh), \end{aligned}$$

where $T_{n1}^+ = \frac{1}{n}\sum_{i=1}^n \int_a^b \mathcal{Z}_i^{+\intercal}(t)\mathcal{Z}_i^+(t)w(t)dt$ and $T_{n2}^+ = \frac{1}{n}\sum_{i\neq k}^n \int_a^b \mathcal{Z}_i^{+\intercal}(t)\mathcal{Z}_k^+(t)w(t)dt$. By similar calculation as in the null hypothesis for T_{n1}, we have $\mathbb{E}(T_{n1}^+) = q + \frac{(m-\bar{r})qh}{\bar{r}\mu_{20}} \int_a^b f(t)w(t)dt + mhb_n^2\mu_0 + O(h^2)$ and $\mathrm{Var}(T_{n1}^+) = O(h^2 + 1/n)$.

For T_{n2}^+, we define the U-statistic as follows

$$\mathcal{U}_n = \frac{T_{n2}^+}{(n-1)} = \frac{1}{n(n-1)} \sum_{i\neq k}^{n} \int_a^b \mathcal{Z}_i^{+\intercal}(t)\mathcal{Z}_k^+(t)w(t)dt = \frac{1}{n(n-1)} \sum_{i\neq k}^{n} \mathcal{K}(\mathcal{Z}_i^+, \mathcal{Z}_k^+),$$

where the kernel function $\mathcal{K}$ is the same as in the proof for the null case. It is easy to show $\theta = \mathbb{E}\mathcal{K}(\mathcal{Z}_1^+, \mathcal{Z}_2^+) = mhb_n^2\mu_0$. And the first projection $\mathcal{K}_1(\mathcal{Z}_1^+) = \mathbb{E}\{\mathcal{K}(\mathcal{Z}_1^+, \mathcal{Z}_2^+)|\mathcal{Z}_1^+\} = -b_n\sqrt{mh}\int_a^b \mathcal{Z}_1^{+\intercal}(t)\mathbf{R}^{1/2}(t)\mathbf{d}(t)w(t)dt$ has the variance ζ_1, which can be obtained by

$$\mathbb{E}\mathcal{K}_1^2(\mathcal{Z}_1^+) = b_n^2 mh \int_a^b \int_a^b \mathbf{d}^\intercal(t)\mathbf{R}^{1/2}(t)\mathbb{E}\{\mathcal{Z}_1^+(t)\mathcal{Z}_1^{+\intercal}(s)\}\mathbf{R}^{1/2}(s)\mathbf{d}(s)w(t)w(s)dtds.$$

Therefore, we have $\zeta_1 = b_n^2(mh)^2\mu_1 + O(b_n^2 mh^2)$, where μ_1 is defined in Theorem 3.

We also have $\zeta_2 = \mathrm{Var}\{\mathcal{K}(\mathcal{Z}_1^+, \mathcal{Z}_2^+)\} = (mh)^2 V + O(h + b_n^2 m^2 h^2)$, where

$$V = 2\int_a^b \int_a^b \mathrm{tr}\{\mathbf{\Sigma}_0(s,t)\mathbf{\Sigma}_0(t,s)\} w(t)w(s)dtds.$$

Thus by U-statistic theory, if $\zeta_2 = o(n\zeta_1)$, which is equivalent to $b_n^{-1} = o(\sqrt{n})$, we have $\mathcal{U}_n \sim \mathrm{AN}(\theta, \frac{4\zeta_1}{n})$ provided that the first projection sequence $\{\mathcal{K}_1(\mathcal{Z}_i^+)\}_{i=1}^n$ satisfy the Lyapunov's condition, which can be verified as follows. Since $\mathbb{E}\mathcal{K}_1(\mathcal{Z}_i^+) = \theta$, $\mathrm{Var}\{\mathcal{K}_1(\mathcal{Z}_i^+)\} = \zeta_1$ and $\mathbb{E}\{\mathcal{K}_1(\mathcal{Z}_i^*) - \theta\}^4 \sim b_n^4 (mh)^4$ up to a constant, we have

$$\frac{\sum_{i=1}^n \mathbb{E}\{\mathcal{K}_1(\mathcal{Z}_i^+) - \theta\}^4}{\left\{\sum_{i=1}^n \mathrm{Var}^2\{\mathcal{K}_1(\mathcal{Z}_i^+)\}\right\}^2} \sim \frac{n b_n^4 (mh)^4}{n^2 b_n^4 (mh)^4} = \frac{1}{n} \to 0.$$

Thus if $b_n^{-1} = o(\sqrt{n})$, we have $T_{n2}^+ \sim \mathrm{AN}(nb_n^2 mh\mu_0, 4nb_n^2 (mh)^2 \mu_1)$. Then the conclusion in part (b) holds.

For part (c) with $\eta > 1/8$, under which we choose the bandwidth $h = n^{-\alpha_0}$ with $1/8 < \alpha_0 < 1/2 - 1/\lambda$. In this scenario, we have $mh \to \infty$. Since $b_n = n^{-1/2}$ and $C_{n,\alpha_0,\eta} = n^{1/2}$, we have

$$\mathbf{U}_n^+(t) = \mathbf{G}(t) C_{n,\alpha_0,\eta}^{-1} \sum_{i=1}^n \xi_i(t) - \mathbf{R}^{1/2}(t)\mathbf{d}(t).$$

By Lemma 9 in Section 3.5.2 in Chapter 2, we know $\mathbf{U}_n^+(t)$ asymptotically converges to a Gaussian process $\mathbf{U}^+(t;\eta)$ with mean $-\mathbf{R}^{1/2}(t)\mathbf{d}(t)$ and covariance function $\mathbf{\Sigma}(s,t)$. Thus the limiting distribution of T_n is the same as the distribution of $Z^+ := \int_a^b \mathbf{U}^+(t;\eta)^T \mathbf{U}^+(t;\eta) w(t)dt$. We only need to show the distribution of Z^+. To this end, using the following eigenvalue decomposition for $\mathbf{U}^+(t;\eta)$ [Bal60] $\mathbf{U}^+(t;\eta) = \sum_{k=1}^\infty \xi_k^+ \boldsymbol{\phi}_k(t)$, where $\xi_k^+ = \int_a^b \mathbf{U}^+(t;\eta)^\intercal \boldsymbol{\phi}_k(t) w(t)dt$ are independent ($k = 1, 2, \cdots, \infty$) normal with mean $-u_k$ and variance γ_k. Here γ_k and

$\boldsymbol{\phi}_k(t)$ is the k-th ordered eigenvalue of $\boldsymbol{\Sigma}_0(s,t)$ and corresponding eigenfunctions in $\mathbb{R}^q$. Then we have

$$Z^+ = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\xi_k^+\xi_l^+\int_a^b \boldsymbol{\phi}_k(t)^\intercal\boldsymbol{\phi}_l(t)w(t)dt = \sum_{k=1}^{\infty}\xi_k^{+2}.$$

Because ξ_k^+ are independent $N(-u_k,\gamma_k)$, we have $T_n \xrightarrow{d} \sum_{k=1}^{\infty}\gamma_k\chi^2_{1,k}\Big(\mu_k^2/\gamma_k\Big)$. This completes the proof of part (c). □

3.5.1.4 Proof of Theorem 4

Proof of Theorem 4. Conditional on the data $\mathcal{X}_n = \{Y_{ij}, X_{ij}, t_{ij}\}_{i=1}^n$, the bootstrapped sample was generated according to $Y_{ij}^* = \tilde{\boldsymbol{\beta}}^\intercal(t_{ij})\mathbf{X}_{ij} + \epsilon_{ij}^*$,which can be regarded an analog of the model (2.1.1) with the true coefficient function $\tilde{\boldsymbol{\beta}}(t)$ and ϵ_{ij}^* has mean 0 and covariance $\hat{\Omega}(s,t)$. Let $o_p^*(1)$ and $O_p^*(1)$ be the stochastic order with respect to the conditional probability measure given the original samples.

Based on this bootstrapped sample $\{Y_{ij}^{*(b)}, t_{ij}, \mathbf{X}_{ij} : i = 1,\cdots,n; j = 1,\cdots,m_i\}$, we first estimate the true $\tilde{\boldsymbol{\beta}}(t)$ by local linear smoothing with the estimator $\hat{\boldsymbol{\beta}}^*(t)$, which differs from the original $\hat{\boldsymbol{\beta}}(t)$ only via the error. And then our estimating equation is constructed as following

$$g_i^*\{\boldsymbol{\beta}(t)\} = \frac{1}{m_i}\sum_{j=1}^{m_i}\Big\{Y_{ij}^* - \boldsymbol{\beta}^\intercal(t)\mathbf{X}_{ij} - \{\hat{\boldsymbol{\beta}}^*(t_{ij}) - \hat{\boldsymbol{\beta}}^*(t)\}^\intercal\mathbf{X}_{ij}\Big\}\mathbf{X}_{ij}K_h(t_{ij}-t).$$

Since we have proved the following results in the proof of Lemma 1 in Section 2.6.2,

$$\sup_{t\in[a,b]}\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij}\mathbf{X}_{ij}^\intercal K_h(t_{ij}-t) - \boldsymbol{\Gamma}(t)f(t)\| = O(\delta_n) \quad \text{a.s.},$$

and by the similar proof as Lemma 2 in Section 2.6.2, we have

$$g_i^*\{\tilde{\boldsymbol{\beta}}(t)\} = \xi_i^*(t)\{1 + \tilde{o}_p^*(1)\} + \tilde{O}(h^4) \quad \text{a.s.}$$

where $\xi_i^*(t) = \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{X}_{ij}\epsilon_{ij}^* K_h(t_{ij} - t)$ and here and below, the almost surely convergence holds with respect to the original probability measure, which is true almost surely for all the sample points in the sample space of $\mathcal{X}_n$ when n is sufficient large. Then by the fact that $\sup_t \|\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\| = O(\delta_{n1} + h^4)$ a.s.. Thus, similar to (2.6.29), we have the following results almost surely

$$\begin{aligned}
2\ell^*(t) &= \frac{1}{C_{n,\alpha_0,\eta}^2}\{\sum_{i=1}^n g_i^*(\tilde{\boldsymbol{\beta}})\}^\intercal \mathbf{A}^{-1}\mathbf{C}^\intercal\mathbf{R}\mathbf{C}\mathbf{A}^{-1}\{\sum_{i=1}^n g_i^*(\tilde{\boldsymbol{\beta}})\} + o_p^*(1) + \tilde{O}(\delta_{n1} + h^4) \\
&= \frac{1}{C_{n,\alpha_0,\eta}^2}\{\sum_{i=1}^n g_i^*(\tilde{\boldsymbol{\beta}})\}^\intercal \mathbf{G}^\intercal\mathbf{G}\{\sum_{i=1}^n g_i^*(\tilde{\boldsymbol{\beta}})\} + o_p^*(1) + \tilde{O}(\delta_{n1} + h^4) \\
&= \mathbf{U}_n^{*\intercal}(t)\mathbf{U}_n^*(t)\{1 + o_p^*(1)\} + \tilde{O}(\delta_{n1} + h^4),
\end{aligned}$$

where $\mathbf{U}_n^*(t) = C_{n,\alpha_0,\eta}^{-1}\mathbf{G}(t)\sum_{i=1}^n \xi_i^*(t)$ with $\mathbf{G}(t) = \mathbf{R}^{1/2}(t)\mathbf{C}(t)\mathbf{A}^{-1}(t)$.

Thus the bootstrapped version test statistic T_n^* can be represented as

$$T_n^* = \int_a^b \mathbf{U}_n^{*\intercal}(t)\mathbf{U}_n^*(t)w(t)dt\{1 + o_p^*(1)\} + o(1) \text{ a.s.} \tag{3.5.6}$$

Let $d(F, G)$ be the maximum norm distance between two distribution functions F and G such that $d(F, G) = \sup_x |F(x) - G(x)|$. From the proof of Theorem 2, we know that the required conditions for showing the convergence of $d(\mathscr{L}(\int_a^b \mathbf{U}_n^\intercal(t)\mathbf{U}_n(t)w(t)dt), \mathscr{L}(T_n)) \to 0$ are the independence between $\mathbf{X}_i(t)$ and $\epsilon_i(t)$, $\epsilon_i(t)$ are independent with $\mathrm{E}\{\epsilon_i(t)\} = 0$ and finite λ moments for $i = 1, \cdots, n$.

To show that $d(\mathscr{L}(\int_a^b \mathbf{U}_n^{*\intercal}(t)\mathbf{U}_n^*(t)w(t)dt|\mathcal{X}_n), \mathscr{L}(T_n)) \to 0, n \to \infty$, we note the difference between $\int_a^b \mathbf{U}_n^{*\intercal}(t)\mathbf{U}_n^*(t)w(t)dt$ and $\int_a^b \mathbf{U}_n^{\intercal}(t)\mathbf{U}_n(t)w(t)dt$ is that $\epsilon_i(t)$ is replaced by $\epsilon_i^*(t)$, which has mean 0 and covariance $\hat{\Omega}(s,t)$. Since $\hat{\Omega}(s,t)$ is a consistent estimator of $\Omega(s,t)$, and from our construction of $\epsilon_i^*(t)$ in the wild bootstrap procedure, given $\mathcal{X}_n$, we have the independence between $\mathbf{X}_i(t)$ and $\epsilon_i^*(t)$, $\mathrm{E}\{\epsilon_i^*(t)\} = 0$ and $\epsilon_i^*(t)$ has finite λ moments. Thus, based on the standard modification of the proof of Theorem 2, we have $d(\mathscr{L}(\int_a^b \mathbf{U}_n^{*\intercal}(t)\mathbf{U}_n^*(t)w(t)dt|\mathcal{X}_n), \mathscr{L}(T_n)) \to 0$. This together with (3.5.6), we have $d(\mathscr{L}(T_n^*|\mathcal{X}_n), \mathscr{L}(T_n)) \to 0$ almost surely. □

3.5.2 Proofs of Proposition and Lemma

Lemma 9. *Under assumptions (C1)-(C4), for the dense functional data, $\mathbf{U}_n(t)$ converges to a multivariate Gaussian process $\xi(t)$ with mean 0 and covariance matrix $\mathbf{\Sigma}_0$ defined in Proposition 4 in Section 3.2.*

Proof. It is clear that $\mathbb{E}\{\mathbf{U}_n(t)\} = C_{n,\alpha_0,\eta}^{-1} \sum_{i=1}^n \mathbf{G}(t)\mathbb{E}\{\xi_i(t)\} = 0$ and

$$\mathrm{Cov}\{\mathbf{U}_n(s), \mathbf{U}_n(t)\} = \frac{1}{C_{n,\alpha_0,\eta}^2} \left\{ \sum_{i=1}^n \mathbf{G}(s)\mathbb{E}\{\xi_i(s)\xi_i^{\intercal}(t)\}\mathbf{G}^{\intercal}(t) \right\}.$$

For computing $\mathbb{E}\{\xi_i(s)\xi_i^{\intercal}(t)\}$, by similar calculation as before, we have the following result,

$$\mathbb{E}\{\xi_i(s)\xi_i^{\intercal}(t)\} = \frac{K^{(2)}(\frac{s-t}{h})}{m_i h}\mathbf{\Gamma}(s)\Omega(s)f(s) + \frac{m_i - 1}{m_i}\mathbf{\Gamma}(s,t)\Omega(s,t)f(s)f(t) + \tilde{O}(h^2),$$

where $K^{(2)}(x) = \int K(y)K(y-x)dy$ and $\mathbf{\Gamma}(s,t) = \mathbb{E}\{\mathbf{X}(s)\mathbf{X}^{\intercal}(t)\}, \Omega(s,t) = \mathbb{E}\{\epsilon(s)\epsilon(t)\}$.

Then we have

$$\begin{aligned}\mathrm{Cov}\{\mathbf{U}_n(s),\mathbf{U}_n(t)\} &= \mathbf{G}(s)\mathbf{\Gamma}(s)\Omega(s)\mathbf{G}^{\intercal}(t)f(s)K^{(2)}(\frac{s-t}{h})\frac{1}{C^2_{n,\alpha_0,\eta}}\sum_{i=1}^{n}\frac{1}{m_i h}\\ &+ \mathbf{G}(s)\mathbf{\Gamma}(s,t)\Omega(s,t)\mathbf{G}^{\intercal}(t)f(s)f(t)\frac{1}{C^2_{n,\alpha_0,\eta}}\sum_{i=1}^{n}\frac{m_i-1}{m_i} \qquad (3.5.7)\\ &+ \frac{n\tilde{O}(h^2)}{C^2_{n,\alpha_0,\eta}}\mathbf{G}(s)\mathbf{G}^{\intercal}(t).\end{aligned}$$

By the definition of $C_{n,\alpha_0,\eta}$, we have the following result,

$$\mathrm{Cov}\{\mathbf{U}_n(s),\mathbf{U}_n(t)\} \sim \mathbf{G}(s)\mathbf{\Gamma}(s,t)\Omega(s,t)\mathbf{G}^{\intercal}(t)f(s)f(t) = \mathbf{\Sigma}_0(s,t).$$

Thus we have $\mathrm{Cov}\{\mathbf{U}_n(s),\mathbf{U}_n(t)\} = \mathbf{\Sigma}_0(s,t)+\tilde{o}(1)$. The proof in Lemma 3 proves the central limit theorem the joint distribution of $\{\mathbf{U}_n(t_1),\cdots,\mathbf{U}_n(t_s)\}$ at finite time points $\{t_1,\cdots,t_s\}$. Weak convergence of $\mathbf{U}_n(t)$ now follows (Billingsley (1968), page 95) if $\forall \mathbf{a}\in\mathbb{R}^q$,

$$\mathbf{a}^{\intercal}\mathbb{E}\left\{[\mathbf{U}_n(s)-\mathbf{U}_n(t)][\mathbf{U}_n(s)-\mathbf{U}_n(t)]^{\intercal}\right\}\mathbf{a} \le C(s-t)^2$$

can be established. To this end, we note that

$$\begin{aligned}&\mathbf{a}^{\intercal}\mathbb{E}\left\{[\mathbf{U}_n(s)-\mathbf{U}_n(t)][\mathbf{U}_n(s)-\mathbf{U}_n(t)]^{\intercal}\right\}\mathbf{a}\\ =&\mathbf{a}^{\intercal}\mathbf{G}(s)\mathbf{B}(s)\mathbf{G}^{\intercal}(s)\mathbf{a}-\mathbf{a}^{\intercal}\mathbf{\Sigma}_0(s,t)\mathbf{a}-\mathbf{a}^{\intercal}\mathbf{\Sigma}_0(t,s)\mathbf{a}+\mathbf{a}^{\intercal}\mathbf{G}(t)\mathbf{B}(t)\mathbf{G}^{\intercal}(t)\mathbf{a}\\ =&2\mathbf{a}^{\intercal}\mathbf{a}-\{\mathbf{a}^{\intercal}[\mathbf{\Sigma}_0(s,s)+(t-s)\frac{\partial\mathbf{\Sigma}_0(s,s)}{\partial t}+(t-s)^2\frac{\partial^2\mathbf{\Sigma}_0(s,s^*)}{\partial t^2}]\mathbf{a}\\ &-\{\mathbf{a}^{\intercal}[\mathbf{\Sigma}_0(t,t)+(s-t)\frac{\partial\mathbf{\Sigma}_0(t,t)}{\partial s}+(t-s)^2\frac{\partial^2\mathbf{\Sigma}_0(t,t^*)}{\partial s^2}]\mathbf{a}\}\\ \le&|s-t||\mathbf{a}^{\intercal}\{\frac{\partial\mathbf{\Sigma}_0(s,s)}{\partial t}-\frac{\partial\mathbf{\Sigma}_0(t,t)}{\partial s}\}\mathbf{a}|+C_1(s-t)^2\le C(s-t)^2,\end{aligned}$$

where we used $\mathbf{\Sigma}_0(s,s) = \mathbf{\Sigma}_0(t,t) = \mathbf{I}_q$ and the last two inequalities follow from the continuity condition (C3). □

Proof of Proposition 4. By (3.5.7) in the proof of Lemma 9, and the definition of $C_{n,\alpha_0,\eta}$, we have the following result, up to a factor $1 + o_p(1)$, $\mathrm{Cov}\{\mathbf{U}_n(s), \mathbf{U}_n(t)\} = \mu_{20}^{-1} K^{(2)}(\frac{s-t}{h})\mathbf{I}_q + mh\mathbf{\Sigma}_0(s,t)$ for $mh \to 0$, and $\mathrm{Cov}\{\mathbf{U}_n(s), \mathbf{U}_n(t)\} = \mathbf{\Sigma}_0(s,t)$ for $mh \to \infty$.

Since $K^{(2)}(\frac{s-t}{h}) = \mu_{20}$ when $s = t$, we can further have, up to a factor $1 + o_p(1)$,

$$\mathrm{Cov}\{\mathbf{U}_n(s), \mathbf{U}_n(t)\} = \begin{cases} \mu_{20}^{-1} K^{(2)}(\frac{s-t}{h})\mathbf{I}_q, & \text{if } m^2h \to 0, \\ \mathbf{I}_q I(s=t) + mh\mathbf{\Sigma}_0(s,t)I(s \neq t) & \text{if } m^2h \to \infty \text{ and } mh \to 0, \\ \mathbf{\Sigma}_0(s,t), & \text{if } mh \to \infty, \end{cases}$$

which complete the proof of the proposition. □

Chapter 4

Empirical Likelihood in Testing Coefficients in High Dimensional Heteroscedastic Linear Models

4.1 Introduction

As mentioned in Section 1.2.2, people have made significant progress towards understanding the estimation theory, but very little work has been done for statistical inference for high dimensional linear models, especially with heteroscedastic noise. Empirical likelihood has the ability of internal studentizing to avoid variance estimation, which can help solve the heteroscedasticity issue.

In Section 4.2, we study the asymptotic normality of Wald type statistic for the existing methods under the heteroscedastic noise. In Section 4.3, we propose the general empirical likelihood framework for analyzing the estimating equations proposed in different ways, although they all follow the low dimensional projection idea. In Section 4.4, we provide implications of the general results on three different cases, projection via lasso estimation, projection via inverse regression and projection via KFC set selection. Section 4.5 provides numerical results and Section 4.6 shows some real data analysis. We refer all of the proofs

to the Technical Details 4.7.

The following notation is adopted throughout this chapter. For $\mathbf{v} = (v_1, v_2, \cdots, v_d)^\intercal \in \mathbb{R}^d$, we define $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$ for $0 < q < \infty$, $\|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})|$ where $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A, and $\|\mathbf{v}\|_\infty = \max_{1 \leq j \leq d} |v_i|$. For a symmetric matrix $\mathbf{M} = ((M_{jk}))$, $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. For any matrix $\mathbf{M} = ((M_{jk}))$, let $\|\mathbf{M}\|_{\max} = \max_{j,k} |M_{jk}|$, $\|\mathbf{M}\|_1 = \max_k \sum_j |M_{jk}|$, $\|\mathbf{M}\|_2 = \sqrt{\lambda_{\max}(\mathbf{M}^\intercal \mathbf{M})}$, and $\|\mathbf{M}\|_\infty = \max_j \sum_k |M_{jk}|$. We denote $\mathbf{I}_d$ as the $d \times d$ identity matrix, and if the dimension is obvious from the context, we just omit the subscript d. For $\mathcal{S} \subseteq \{1, 2 \cdots, d\}$, let $\mathbf{v}_\mathcal{S} = \{v_j : j \in \mathcal{S}\}$ be a subvector of $\mathbf{v}$. And for any $k \in \{1, 2, \cdots, d\}$, let $\mathbf{M}_{j\mathcal{S}} = \{M_{jl}, l \in \mathcal{S}\}$ as a row vector and $\mathbf{M}_{\mathcal{S}j} = \{M_{lj} : l \in \mathcal{S}\}$ as a column vector. Denote $\backslash k = \{1, 2, \cdots, k-1, k+1, \cdots, d\}$. For a sequence of random variables X_n, we write $X_n \xrightarrow{d} X$ for some random variable X, if X_n converges to X in distribution, and write $X_n \xrightarrow{p} a$ for some constant a, if X_n converges in probability to a. For notational simplicity, we use $C, C', C'', C_1, C_2, C_3$ to denote generic constants, whose values can change from line to line.

4.2 Preliminary and Existing Methods

We consider a linear regression model:

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \tag{4.2.1}$$

where $\mathbb{Y} = (Y_1, Y_2, \cdots, Y_n)^\intercal \in \mathbb{R}^n$ is a response vector, $\mathbb{X} = ((X_{ij})) \in \mathbb{R}^{n \times p}$ is a random design matrix with columns $\{\mathbb{X}_j \in \mathbb{R}^n\}_{j=1}^p$ and rows $\{\mathbf{X}_i \in \mathbb{R}^p\}_{i=1}^n$, which are assumed

to be independent and identically distributed (IID) with $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{X}_i) = \mathbf{\Sigma}$, and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is a vector of unknown true regression coefficients. The error term satisfies $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0$, and $\mathrm{Var}(\epsilon_i|\mathbf{X}_i) = \sigma_\epsilon^2(\mathbf{X}_i)$, which allows heteroscedasticity. Note that with these assumptions, $\mathbf{X}_i$ and ϵ_i are uncorrelated, i.e. $\mathrm{E}(\mathbf{X}_i\epsilon) = \mathbf{0}$. In addition, we assume the marginal variance $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$. Hereafter we assume that $p \gg n$. Denote $s = \|\boldsymbol{\beta}^0\|_0$ be the number of non-zeros of $\boldsymbol{\beta}^0$ and we assume sparsity with $s < n$. Let $\mathbf{Z}_i = \epsilon_i\mathbf{X}_i$ be a random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Theta} = ((\theta_{jk}))$. And assume $\mathrm{Var}(\epsilon_i^2) = \kappa$ and $\mathrm{Cov}(\epsilon_i^2, \mathbf{Z}_i) = \boldsymbol{\varpi}$.

In practice, among hundreds of thousands of regressors, people want to test whether some target features are significant or not. For example, one may want to know whether a particular gene effect is significant or not among thousands of genes. To assess the significance of a single coefficient, we test the following hypothesis for any given $j \in \{1, 2, \cdots, p\}$,

$$H_0 : \beta_j^0 = 0 \quad vs. \quad H_1 : \beta_j^0 \neq 0. \tag{4.2.2}$$

Statistical inference for low-dimensional coefficients in high dimensional linear model with homoscedastic noise has received increasing attention. Low dimensional projection method has been introduced by [ZZ14] and [B$^+$13].

Under (4.2.1), and in low dimensional scenario, i.e. $p \leq n$, we have the ordinary least square (OLS) estimator for β_j^0,

$$\hat{\beta}_j = \frac{(\mathbb{X}_j^\perp)^\intercal \mathbb{Y}}{(\mathbb{X}_j^\perp)^\intercal \mathbb{X}_j} = \frac{(\mathcal{Q}_{\setminus j}\mathbb{X}_j)^\intercal \mathbb{Y}}{(\mathcal{Q}_{\setminus j}\mathbb{X}_j)^\intercal \mathbb{X}_j} = \frac{(\mathcal{Q}_{\setminus j}\mathbb{X}_j)^\intercal (\mathcal{Q}_{\setminus j}\mathbb{Y})}{(\mathcal{Q}_{\setminus j}\mathbb{X}_j)^\intercal (\mathcal{Q}_{\setminus j}\mathbb{X}_j)} = \frac{\mathbb{X}_j^\intercal \mathcal{Q}_{\setminus j}\mathbb{Y}}{\mathbb{X}_j^\intercal \mathcal{Q}_{\setminus j}\mathbb{X}_j}, \tag{4.2.3}$$

where $\mathbb{X}_j^\perp$ is the projection of $\mathbb{X}_j$ to the orthogonal complement of the column space spaned

by $\{\mathbb{X}_{\setminus j}\}$, and $\mathcal{Q}_{\setminus j}$ is as defined below for general $\mathcal{Q}_{\mathcal{S}}$ with $\mathcal{S} \subseteq \{1, 2 \cdots, p\}$ and $|\mathcal{S}| < n$,

$$\mathcal{Q}_{\mathcal{S}} = \mathbf{I} - \mathcal{P}_{\mathcal{S}} = \mathbf{I} - \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal} \in \mathbb{R}^{n \times n}.$$

However in the high dimensional linear model with $p > n$, the OLS estimator is no longer valid. Instead of projection onto the space spanned by all of the rest covariates, people select the projection space based on the correlations between X_j and the others.

4.2.1 Lasso Projection

In [ZZ14, vdGBR13, NL14], they used the linear sparse regularized regression procedure such as Lasso to select the projection space. Define $\eta_{ij} := X_{ij} - \mathbf{X}_{i,\setminus j}^{\intercal}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j}$. That is

$$X_{ij} = \mathbf{X}_{i,\setminus j}^{\intercal}\mathbf{w}_j^0 + \eta_{ij}$$

with $\mathbf{w}_j^0 = \boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j}$, which leads to the following generalized version of (4.2.3) with relaxed projection

$$\hat{\beta}_j^{(\mathrm{lin})} = \frac{\mathbb{Z}_j^{\intercal}\mathbb{Y}}{\mathbb{Z}_j^{\intercal}\mathbb{X}_j}, \text{where } \mathbb{Z}_j = \mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\mathbf{w}}_j \tag{4.2.4}$$

with $\hat{\mathbf{w}}_j$ as an estimator of $\mathbf{w}_j^0$. However, $\hat{\beta}_j^{(\mathrm{lin})}$ is biased. To solve this issue, [ZZ14] proposed the de-biased estimator as follows,

$$\hat{\beta}_j^{(\mathrm{de})} = \frac{\mathbb{Z}_j^{\intercal}\mathbb{Y} - \sum_{k \neq j}\mathbb{Z}_j^{\intercal}\mathbb{X}_k\hat{\beta}_k}{\mathbb{Z}_j^{\intercal}\mathbb{X}_j}, \tag{4.2.5}$$

where $\hat{\boldsymbol{\beta}}$ is some initial estimator of $\boldsymbol{\beta}^0$. This de-biased estimator (4.2.5) can be regarded as the solution to the estimating equation, which is based on the population subject $\eta_{ij}\epsilon_i = \{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i,\setminus j})\}\{Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta}^0\}$, that is

$$\sum_{i=1}^{n} m_{ni}^{(\mathrm{lasso})}(\beta_j) := \sum_{i=1}^{n} \{X_{ij} - \mathbf{X}_{i,\setminus j}^\intercal \hat{\mathbf{w}}_j\}\{Y_i - X_{ij}\beta_j - \mathbf{X}_{i,\setminus j}^\intercal \hat{\boldsymbol{\beta}}_{\setminus j}\} = 0. \tag{4.2.6}$$

And by simple algebra, we have

$$m_{ni}^{(\mathrm{lasso})}(\beta_j^0) = \underbrace{\epsilon_i \eta_{ij}}_{W_{ni}^{(\mathrm{lasso})}} + \underbrace{\eta_{ij}(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})^\intercal \mathbf{X}_{i,\setminus j} + (\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal \mathbf{X}_{i,\setminus j}\{Y_i - X_{ij}\beta_j^0 - \mathbf{X}_{i\setminus j}\hat{\boldsymbol{\beta}}_{\setminus j}\}}_{R_{ni}^{(\mathrm{lasso})}}.$$

By simple calculation, we have $\mathrm{E}(W_{ni}^{(\mathrm{lasso})}) = \mathrm{E}\{\epsilon_i(X_{ij} - \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{X}_{i,\setminus j})\} = 0$ and

$$\begin{aligned}
\mathrm{E}[(W_{ni}^{(\mathrm{lasso})})^2] &= \mathrm{E}\{\epsilon_i^2(X_{ij} - \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{X}_{i,\setminus j})^2\} \\
&= \mathrm{E}\{\epsilon_i^2(X_{ij}^2 - 2X_{ij}\boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{X}_{i,\setminus j} + \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{X}_{i,\setminus j}\mathbf{X}_{i,\setminus j}^\intercal\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j})\} \\
&= \mathrm{E}\{Z_{ij}^2 - 2Z_{ij}\boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{Z}_{i,\setminus j} + \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\mathbf{Z}_{i,\setminus j}\mathbf{Z}_{i,\setminus j}^\intercal\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j}\} \\
&= \theta_{jj} - 2\boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Theta}_{j,\setminus j} + \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Theta}_{\setminus j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j} := \sigma_{n,\mathrm{lasso}}^2.
\end{aligned}$$

Note that if we assume the independence between the error term and the covariates, we have the following simplified form

$$\mathrm{E}[(W_{ni}^{(\mathrm{lasso})})^2] = \sigma_\epsilon^2(\sigma_{jj} - \boldsymbol{\Sigma}_{j,\setminus j}\boldsymbol{\Sigma}_{\setminus j,\setminus j}^{-1}\boldsymbol{\Sigma}_{\setminus j,j}).$$

This shows the difference between our heteroscedastic case and the homoscedastic case.

For the homoscedastic case, as discussed in [ZZ14] [vdGBR13], the inference proce-

dure based on asymptotic normality needs to estimate the asymptotic variance $\sigma_\epsilon^2/(\sigma_{jj} - \mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Sigma}_{\backslash j,j})$. Under the heteroscedastic noise, we can still show the following asymptotic normality but with much more complicated asymptotic variance.

Proposition 5. *Under model (4.2.1) with heteroscedastic noise, if Assumption 1 in the appendix holds, we have*

$$\sqrt{n}(\hat{\beta}_j^{(de)} - \beta_j^0) \xrightarrow{d} N(0, \sigma_{lasso}^2) \tag{4.2.7}$$

where the asymptotic variance is defined as follows

$$\sigma_{lasso}^2 = \lim_{n\to\infty} \frac{\theta_{jj} - 2\mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Theta}_{j,\backslash j} + \mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Theta}_{\backslash j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Sigma}_{\backslash j,j}}{(\sigma_{jj} - \mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Sigma}_{\backslash j,j})^2}. \tag{4.2.8}$$

Such complex asymptotic variance (4.2.8) makes it hard to use Wald type inference procedure in practice since it is difficulty to get a good estimate for the asymptotic variance. Thus naively using the Wald type test procedure proposed by [ZZ14] in the heteroscedastic case will lead to invalid results, which will be demonstrated in the simulation study in Section 4.5.

4.2.2 KFC Projection

[LZL+13] proposed another way to select the projection space, which is based on the so called KFC set $\mathcal{S} = \{l \neq j : |\sigma_{jl}| > c\}$ for some pre-specified threshold value $c > 0$. That is essentially the set of all key confounders associated with X_j. And then the estimator can

be obtained by the projection with respect to the covariates indexed by $\mathcal{S}$,

$$\hat{\beta}_j^{(\text{kfc})} = \frac{\mathbb{X}_j^\intercal \mathcal{Q}_\mathcal{S} \mathbb{Y}}{\mathbb{X}_j^\intercal \mathcal{Q}_\mathcal{S} \mathbb{X}_j} = \frac{\tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{Y}}}{\tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{X}}_j}, \tag{4.2.9}$$

with the profiled response and target predictor as $\tilde{\mathbb{Y}} = \mathcal{Q}_\mathcal{S}\mathbb{Y}, \quad \tilde{\mathbb{X}}_j = \mathcal{Q}_\mathcal{S}\mathbb{X}_j$.

Based on the de-bias idea, we propose the following de-biased KFC estimator

$$\hat{\beta}_j^{(\text{kfc-de})} = \frac{\tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{Y}} - \sum_{k\in\mathcal{S}^*} \tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{X}}_k \hat{\beta}_k}{\tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{X}}_j}, \tag{4.2.10}$$

where $\mathcal{S}^* = \mathcal{S}^{+c}$, i.e. the complement of $\mathcal{S}^+ := \{1\} \cup \mathcal{S}$, and $\hat{\boldsymbol{\beta}}_{\mathcal{S}^*}$ is an initial estimator.

In fact, the above de-biased KFC estimator is the solution to the estimating equation based on the population subject $\eta_{ij,\mathcal{S}}\epsilon_i := \big\{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i\mathcal{S}})\big\}\big\{Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta}^0\big\}$, that is

$$\sum_{i=1}^n m_{ni}^{(\text{kfc})}(\beta_j) := \sum_{i=1}^n (\tilde{Y}_i - \tilde{X}_{ij}\beta_j - \tilde{\mathbf{X}}_{i\mathcal{S}^*}^\intercal \hat{\boldsymbol{\beta}}_{\mathcal{S}^*})\tilde{X}_{ij} = 0, \tag{4.2.11}$$

where $m_n^{(\text{kfc})}(\beta_j^0)$ can be decomposed as

$$\begin{aligned} m_{ni}^{(\text{kfc})}(\beta_j^0) = {} & \epsilon_i \eta_{ij,\mathcal{S}} + \big\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} - X_{ij}\big\}\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_\mathcal{S})^{-1}\mathbb{X}_\mathcal{S}^\intercal\boldsymbol{\epsilon} \\ & + \big\{\epsilon_i - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_\mathcal{S})^{-1}\mathbb{X}_\mathcal{S}^\intercal\boldsymbol{\epsilon}\big\}\big\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^\intercal\mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_\mathcal{S})^{-1}\mathbf{X}_{i\mathcal{S}}\big\} \\ & + \big\{X_{ij} - \mathbb{X}_j^\intercal\mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_\mathcal{S})^{-1}\mathbf{X}_{i\mathcal{S}}\big\}\big\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_\mathcal{S})^{-1}\mathbb{X}_\mathcal{S}^\intercal\mathbb{X}_{\mathcal{S}^*}\big\}[\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}]. \end{aligned}$$

We denote the first term as $W_{ni}^{(\text{kfc})}$ and all the others are denoted by $R_{ni}^{(\text{kfc})}$. And for simplicity we assume the normality of $\mathbf{X}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for the KFC projection section. Now

$W_{ni}^{(\text{kfc})} = \{\epsilon_i(X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}})\}_{i=1}^n$ are IID with $\mathrm{E}W_{ni}^{(\text{kfc})} = 0$ and

$$\begin{aligned}
\mathrm{E}[(W_{ni}^{(\text{kfc})})^2] &= \mathrm{E}\big\{\epsilon_i^2(X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}})^2\big\} \\
&= \mathrm{E}\{\epsilon_i^2(X_{ij}^2 - 2X_{ij}\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} + \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})\} \\
&= \mathrm{E}\{Z_{ij}^2 - 2Z_{ij}\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{Z}_{i\mathcal{S}} + \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{Z}_{i\mathcal{S}}\mathbf{Z}_{i\mathcal{S}}^\intercal\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j}\} \\
&= \theta_{jj} - 2\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{j\mathcal{S}} + \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{\mathcal{S}\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j}.
\end{aligned}$$

Note that if we assume independence between ϵ_i and $\mathbf{X}_i$, we have $\mathrm{E}[(W_{ni}^{(\text{kfc})})^2] = \sigma_\epsilon^2(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})$.

Thus if we assume independence between ϵ_i and $\mathbf{X}_i$, we have the simple asymptotic variance for $\hat{\beta}_j^{(\text{kfc-de})}$, $\sigma_\epsilon^2/(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})$ as discussed in [LZL+13]. But under model (4.2.1) with heteroscedastic error term, we have the following asymptotic normality with more complicated variance.

Proposition 6. *Under the Assumption 3 in the appendix, we have*

$$\sqrt{n}(\hat{\beta}_j^{(kfc\text{-}de)} - \beta_j^0) \xrightarrow{d} N(0, \sigma_{kfc}^2), \tag{4.2.12}$$

where the asymptotic variance is defined as

$$\sigma_{kfc}^2 = \lim_{n\to\infty}(\theta_{jj} - 2\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{j\mathcal{S}} + \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{\mathcal{S}\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})/(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})^2. \tag{4.2.13}$$

Similarly, since the expression (4.2.13) for the asymptotic variance is really complicated, which makes such Wald type statistic hard to use in practice.

4.2.3 Inverse Projection

So far we construct estimators for the target coefficient parameter β_j directly. However, to conduct the hypothesis testing problem (4.2.2), [LL14] proposed an equivalent test based on the projection of X_{ij} onto $(Y_i, \mathbf{X}_{i,\backslash j}^\intercal)^\intercal$,

$$X_{ij} = (Y_i, \mathbf{X}_{i,\backslash j}^\intercal)\boldsymbol{\gamma}_j^0 + \eta_{ij,y}, \tag{4.2.14}$$

where $\eta_{ij,y}$ satisfies $\mathrm{E}\eta_{ij,y} = 0, \mathrm{Cov}(\eta_{ij,y}, (Y_i, \mathbf{X}_{i,\backslash j}^\intercal)) = \mathbf{0}$. Under the linear model (4.2.1) with heteroscedastic noise, as long as $\mathrm{Cov}(\mathbf{X}_i, \epsilon) = \mathbf{0}$, we can still show that the vector $\boldsymbol{\gamma}_j^0$ satisfies $\boldsymbol{\gamma}_j^0 = -\sigma_{\eta_{j,y}}^2\big(-\frac{\beta_j^0}{\sigma_\epsilon^2}, \frac{\beta_j^0\boldsymbol{\beta}_{\backslash j}^{0\intercal}}{\sigma_\epsilon^2} + \boldsymbol{\Omega}_{\backslash j,j}\big)^\intercal$, where $\sigma_{\eta_{j,y}}^2 = \mathrm{Var}(\eta_{ij,y}) = ((\beta_j^0)^2 + w_{jj})^{-1}$ with $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = ((w_{jk}))$. Because $\mathrm{Cov}(\epsilon_i, \mathbf{X}_i) = \mathbf{0}$, we have

$$\mathrm{Cov}(\epsilon_i, \eta_{ij,y}) = \gamma_{j1}^0\mathrm{Cov}(\epsilon_i, -Y_i) = -\sigma_{\eta_{j,y}}^2\beta_j^0 := -\mathfrak{b}_j^0. \tag{4.2.15}$$

Hence the test (4.2.2) is equivalent to $H_0 : \mathfrak{b}_j^0 = 0$. Based on the idea proposed in [LL14], we can have the estimation for $\mathfrak{b}_j^0$

$$\hat{\mathfrak{b}}_j = -\frac{1}{n}\sum_{i=1}^n \big\{Y_i - \mathbf{X}_i^\intercal\hat{\boldsymbol{\beta}}\big\}\big\{X_{ij} - (Y_i, \mathbf{X}_{i,\backslash j}^\intercal)\hat{\boldsymbol{\gamma}}_j\big\} \tag{4.2.16}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_j$ are some initial estimators for $\boldsymbol{\beta}^0$ and $\boldsymbol{\gamma}_j^0$.

Observe that $\hat{\mathfrak{b}}_j$ is the solution to the estimating equation based on $\eta_{ij,y}\epsilon_i + \mathfrak{b}_j^0 = \big\{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i,\backslash j}, Y_i)\big\}\big\{Y_i - \mathbf{X}_i^\intercal\boldsymbol{\beta}^0\big\} + \sigma_{\eta_j,y}^2\beta_j^0$, that is

$$\sum_{i=1}^n m_{ni}^{(\mathrm{inv})}(\mathfrak{b}_j) := \sum_{i=1}^n \big\{Y_i - \mathbf{X}_i^\intercal\hat{\boldsymbol{\beta}}\big\}\big\{X_{ij} - (Y_i, \mathbf{X}_{i,\backslash j}^\intercal)\hat{\boldsymbol{\gamma}}_j\big\} + n\mathfrak{b}_j = 0, \tag{4.2.17}$$

and also by simple algebra, we have

$$m_{ni}^{(\text{inv})}(\mathfrak{b}_j^0) = \underbrace{\{\epsilon_i \eta_{ij,y} + \mathfrak{b}_j^0\}}_{W_{ni}^{(\text{inv})}} + \underbrace{\epsilon_i (Y_i, \mathbf{X}_{i,\setminus j}^\intercal)(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j) + \mathbf{X}_i^\intercal (\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\big\{X_{ij} - (Y_i, \mathbf{X}_{i,\setminus j}^\intercal)\hat{\boldsymbol{\gamma}}_j\big\}}_{R_{ni}^{(\text{inv})}}.$$

With simple calculations, we have $\mathrm{E}(W_{ni}) = 0$ and

$$\begin{aligned}
\mathrm{Var}(W_{ni}) &= \mathrm{Var}(\epsilon_i \eta_{ij,y}) = \mathrm{Var}(\epsilon_i (X_{ij} - \mathbf{X}_i^\intercal \boldsymbol{\beta}^0 \gamma_{j1}^0 - \epsilon_i \gamma_{j1}^0 - \mathbf{X}_{i,\setminus j}^\intercal \boldsymbol{\gamma}_{j,\setminus 1}^0)) \\
&= \theta_{jj} + (\gamma_{j1}^0)^2 \boldsymbol{\beta}^{0\intercal} \boldsymbol{\Theta} \boldsymbol{\beta}^0 + (\gamma_{j1}^0)^2 \kappa + \boldsymbol{\gamma}_{j,\setminus 1}^{0\intercal} \boldsymbol{\Theta}_{\setminus j,\setminus j} \boldsymbol{\gamma}_{j,\setminus 1}^0 \\
&\quad - 2\gamma_{j1}^0 \boldsymbol{\beta}^{0\intercal} \boldsymbol{\Theta}_{\cdot,j} - 2\gamma_{j1}^0 \varpi_j - 2\boldsymbol{\gamma}_{j,\setminus 1}^{0\intercal} \boldsymbol{\Theta}_{\setminus j,j} + 2(\gamma_{j1}^0)^2 \boldsymbol{\beta}^{0\intercal} \boldsymbol{\varpi} \\
&\quad + 2\gamma_{j1}^0 \boldsymbol{\beta}^{0\intercal} \boldsymbol{\Theta}_{\cdot,\setminus j} \boldsymbol{\gamma}_{j,\setminus 1}^0 + 2\gamma_{j1}^0 \boldsymbol{\gamma}_{j,\setminus 1}^{0\intercal} \boldsymbol{\varpi}_{\setminus j} := \sigma_{n,\text{inv}}^2.
\end{aligned}$$

Note that if we assume the independence between ϵ_i and $\mathbf{X}_i$, we have the following simplified variance expression. Since $X_{ij} = \mathbf{X}_i^\intercal \boldsymbol{\beta}^0 \gamma_{j1}^0 + \epsilon_i \gamma_{j1}^0 + \mathbf{X}_{i,\setminus j}^\intercal \boldsymbol{\gamma}_{j,\setminus 1}^0 + \eta_{ij,y}$ and $\mathrm{Cov}(\epsilon_i, \mathbf{X}_i) = 0$, we have $\mathrm{Cov}(\epsilon_i, \epsilon_i \gamma_{j1}^0 + \eta_{ij,y}) = 0$, i.e. $-\gamma_{j1}^0 \mathrm{Var}(\epsilon_i) = \mathrm{Cov}(\epsilon_i, \eta_{ij,y})$. Hence

$$\begin{aligned}
\mathrm{Var}(W_{ni}) &= \mathrm{Var}(\epsilon_i (\eta_{ij,y} + \epsilon_i \gamma_{j1}^0) - \epsilon_i^2 \gamma_{j1}^0) \\
&= \mathrm{Var}(\epsilon_i)\mathrm{Var}(\eta_{ij,y}) + (\gamma_{j1}^0)^2 (\mathrm{Var}(\epsilon_i^2) - \mathrm{Var}^2(\epsilon_i)).
\end{aligned}$$

If furthermore we assume normality for the error term then we have $\mathrm{Var}(\epsilon_i^2) - \mathrm{Var}^2(\epsilon_i) = \mathrm{E}(\epsilon_i^4) - 2[\mathrm{E}(\epsilon_i^2)]^2 = 3\sigma_\epsilon^4 - 2\sigma_\epsilon^4 = \mathrm{Var}^2(\epsilon_i)$, which leads to the same result in Theorem 3.1

from [LL14], i.e.

$$\begin{aligned}\mathrm{Var}(W_{ni}) &= \mathrm{Var}(\epsilon_i)\mathrm{Var}(\eta_{ij,y}) + (\gamma_{j1}^0)^2(\mathrm{Var}(\epsilon_i^2) - \mathrm{Var}^2(\epsilon_i)) \\ &= \mathrm{Var}(\epsilon_i)\mathrm{Var}(\eta_{ij,y}) + [\mathrm{Cov}(\epsilon_i, \eta_{ij,y})]^2 = \sigma_\epsilon^2\sigma_{\eta_j,y}^2 + (\gamma_{j1}^0)^2\sigma_\epsilon^4 \\ &= \sigma_\epsilon^2\sigma_{\eta_j,y}^2 + (\beta_j^0)^2\sigma_{\eta_j,y}^4,\end{aligned}$$

which is more likely to be estimable.

But we can still get the asymptotic normality as stated in the following proposition.

Proposition 7. *Under Assumption 2 in the appendix, we have*

$$\sqrt{n}(\hat{\mathfrak{b}}_j - \mathfrak{b}_j^0) \xrightarrow{d} N(0, \sigma_{inv}^2) \tag{4.2.18}$$

where $\sigma_{inv}^2 = \lim_{n\to\infty} \sigma_{n,inv}^2$.

But we see that the asymptotic variance of $\hat{\mathfrak{b}}_j$ is too way complicated, which makes such Wald type statistics hard to use in practice with heteroscedastic noise.

4.3 Empirical Likelihood Based Approach

To avoid the complexity of estimating asymptotic variance under heteroscedasitic case, we propose EL based approach. Note that the above three procedures in Sections 4.2.1, 4.2.2 and 4.2.3 correspond to three estimating equations (4.2.6), (4.2.11) and (4.2.17) of the form $m_n(\mathbf{X}_i, Y_i, \beta_j, \hat{\boldsymbol{\beta}}_{\setminus j}, \hat{\boldsymbol{\theta}})$, where the nuisance parameters $\boldsymbol{\beta}_{\setminus j}$ and the other nuisance parameters denoted as $\boldsymbol{\theta}$ replaced by their estimators $\hat{\boldsymbol{\beta}}_{\setminus j}$ and $\hat{\boldsymbol{\theta}}$. To keep it simple, we write $m_{ni}(\beta_j) = m_n(\mathbf{X}_i, Y_i, \beta_j, \hat{\boldsymbol{\beta}}_{\setminus j}, \hat{\boldsymbol{\theta}})$ in general.

Note that the estimating equations (4.2.6), (4.2.11) and (4.2.17) have the same structure,

i.e. the first term is the population level term, which will be shown to be dominant and asymptotically normal, while the other terms are all about estimation errors, which need to be controlled. We propose the following general framework by assuming the estimating equations evaluated at the truth β_j^0 can be decomposed as follows,

$$m_{ni}(\beta_j^0) := m_n(\mathbf{X}_i, Y_i, \beta_j^0, \hat{\boldsymbol{\beta}}_{\setminus j}, \hat{\boldsymbol{\theta}}) := W_{ni} + R_{ni} \tag{4.3.19}$$

where $\{W_{ni}\}_{i=1}^n$ which are IID and $\{R_{ni}\}_{i=1}^n$ need to satisfy the following conditions:

(C0) $\mathrm{P}\big\{ \min_{1\le i\le n} m_{ni} < 0 < \max_{1\le i\le n} m_{ni} \big\} \to 1$;

(C1) W_{ni}'s are IID with mean 0 and finite variance σ_n^2 with $\sigma_n^2 \to \sigma_w^2$;

(C2) $\frac{1}{\sqrt{n}} \sum_{i=1}^n R_{ni} = o_p(1)$ and $\max_{1\le i\le n} |R_{ni}| = o_p(n^{1/2})$.

According to [Owe01], with estimating equations, we can construct empirical likelihood to make the inference. Define the following empirical likelihood ratio function of the target parameter β_j

$$\mathrm{EL}_n(\beta_j) = \max \Big\{ \prod_{i=1}^n np_i : p_i > 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i m_{ni}(\beta_j) = 0 \Big\}. \tag{4.3.20}$$

Under this unified framework with the above general conditions, we have the following powerful Wilks theorem.

Theorem 5. *If (C0)-(C2) hold, then*

$$-2 \log EL_n(\beta_j^0) \xrightarrow{d} \chi_1^2.$$

Based on Theorem 5, an asymptotic α level test is given by rejecting H_0 if $-2\log \mathrm{EL}_n(\beta_j^0) > \chi^2_{1,\alpha}$ where $\chi^2_{1,\alpha}$ is the upper α quantile of χ^2_1. We can also construct a $(1-\alpha)100\%$ confidence interval for β_j as $\mathrm{CI}_\alpha = \{\beta_j : -2\log \mathrm{EL}_n(\beta_j) < \chi^2_{1,\alpha}\}$. Since the asymptotic distribution is chi-square, we do not need to estimate any additional parameters, such as the asymptotic variance.

4.4 Theoretical Examples

This section outlines three examples as we discussed above in Sections (4.2.1), (4.2.2) and (4.2.3) to demonstrate interesting and powerful applications of Theorems 5. We need to check the conditions (C0)-(C2) for these problems.

From Proposition 5, 6 and 7, we see that Wald type inference procedure is hard to implement due to the complex asymptotic variance. Fortunately we do not need to estimate that variance in order to conduct inference by using the self studentized EL procedure. And in fact, we already verified condition (C1) for the three procedures in Section (4.2.1), (4.2.2) and (4.2.3), respectively. We can control the second term R_{ni}s under certain assumptions, which leads to the following theorems.

4.4.1 Lasso Projection

The first example is about using Lasso estimation to get the low dimensional projection as we discussed in Section 4.2.1.

Theorem 6. *Under some typical conditions for the initial estimators as in Assumption 1 in the appendix and assume that* $\mathbf{X}_i$ *and* ϵ_i *are both sub-Gaussian. As long as* $s\log p/\sqrt{n} = o(1)$*, the conditions (C0) and (C2) can be satisfied. Assume* $\sigma^2_{n,lasso} \to \sigma^2_{lasso}$ *for some* $\sigma^2_{lasso} < \infty$*,*

and then we have

$$-2\log EL_n^{(lasso)}(\beta_j^0) \xrightarrow{d} \chi_1^2.$$

Notice that under the homoscedastic noise case, [ZZ14] and [vdGBR13] used the Wald type test statistic for testing H_0 based on the same estimation equation as we used here. And in [NL14], with the same estimating equation, they instead proposed the Score test statistic for testing H_0. Although they are asymptotically equivalent, the differences between these two can be found in [NL14]. We are using the same estimating equation to construct the likelihood ratio type statistic for testing H_0. Since we are using empirical likelihood, it not only enjoys the Wilks phenomenon, but also has other nice properties, such as the shape of the confidence interval is data driven and our procedure is more robust to the distribution assumption for the error term since it only requires moment assumptions. The key advantage of our method is that we allow heteroscedasticity for the error term due to the self studentization property of the empirical likelihood. Please refer to the empirical studies in the simulation section for the performance comparison of our method with the Wald type test and Score test.

4.4.2 Inverse Projection

The second example is about using inverse regression to get the low dimensional projection as we discussed in Section 4.2.3.

Theorem 7. *Under some conditions for the initial estimators as in Assumption 2 in the appendix, and assume* $(\mathbf{X}_i^\intercal, \epsilon_i)^\intercal$ *is sub-Gaussian. As long as* $s\log p/\sqrt{n} = o(1)$*, the conditions (C0) and (C2) can be satisfied. Assume* $\sigma_{n,inv}^2 \to \sigma_{inv}^2$ *for some* $\sigma_{inv}^2 < \infty$*, and then*

we have

$$-2\log EL_n^{(inv)}(\mathfrak{b}_j^0) \xrightarrow{d} \chi_1^2.$$

Note that since we are doing an equivalent test, from this inference procedure, we can not get the confidence interval for β_j^0.

4.4.3 KFC Projection

The third example is about the projection by selecting the KFC set as we discussed in Section 4.2.2.

Theorem 8. *Under Assumption 3 in the appendix, the conditions (C0) and (C2) can be satisfied. Assume $\sigma_{n,kfc}^2 \to \sigma_{kfc}^2$ for some $\sigma_{kfc}^2 < \infty$, and then we have*

$$-2\log EL_n^{(kfc)}(\beta_j^0) \xrightarrow{d} \chi_1^2.$$

About the KFC set selection, we propose the following procedure. Based on normality assumption of the predictors, we have the well known conditional distribution result for any give subset $\mathcal{S}$:

$$\rho_{jk}(\mathcal{S}) := \mathrm{Corr}(X_{ij}, X_{ik}|\mathbf{X}_{i\mathcal{S}}) = \sigma_{jk} - \mathbf{\Sigma}_{\mathcal{S}j}^\intercal \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{\Sigma}_{\mathcal{S}k}.$$

The sample partial correlation can be evaluated by, $\hat{\rho}_{jk}(\mathcal{S}) = \tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{X}}_k/n$. For testing whether a partial correlation is zero or not, we could apply Fisher's z-transformation

$$\hat{F}_{jk} = \frac{1}{2}\log\left\{\frac{1+\hat{\rho}_{jk}(\mathcal{S})}{1-\hat{\rho}_{jk}(\mathcal{S})}\right\}.$$

Classical decision theory yields the following rule when using the significance level α. Reject

the null hypothesis $H_0 : \rho_{jk}(\mathcal{S}) = 0$ against the two-sided alternative $H_a : \rho_{jk}(\mathcal{S}) \neq 0$ if

$$\sqrt{n - |\mathcal{S}| - 3}|\hat{F}_{jk}| > z_{1-\alpha/2}.$$

So we could then select the smallest size of $\mathcal{S}$ such that

$$\max_{k \in \mathcal{S}^*} \sqrt{n - |\mathcal{S}| - 3}|\hat{F}_{jk}| < z_{1-\alpha/2}.$$

And in order to make this KFC set selection more stable, we adopt the stability selection proposed by [MB10] and [SS13]. According to [SS13], we split the data into half for B times and select the final KFC set with variables shown at least 50% of those $2B$ KFC sets.

4.5 Simulation Studies

In this section, we conduct simulation studies to investigate the finite sample performance of the proposed empirical likelihood ratio test, as well as comparing the performances for different estimating equations proposed in the existing literature. In particular, to generate the covariates, we simulate $n = 200, 400$ independent samples from a multivariate Gaussian distribution $\mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma})$ where $p = 100, 200, 500$. We consider 3 different covariance matrices $\mathbf{\Sigma} = ((\sigma_{jk}))$, banded matrix with $\sigma_{jk} = \rho^{|j-k|}\mathbb{1}(|j-k| < 2)$, Toeplitz matrix with $\sigma_{jk} = \rho^{|j-k|}$ and block diagonal matrix with unit block $\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$, where $\rho = 0.2, 0.5$. We consider five scenarios for the error distribution, standard normal $\mathrm{N}(0, 1)$, mixture normal distribution $0.7\mathrm{N}(0, 1) + 0.3\mathrm{N}(0, 5^2)$, t distribution with degrees of free-

dom 3, and two heteroscedastic distributions $0.7X_1Z$ and $\frac{1}{p-1}X_1Z\sum_{j=2}^{p}X_{j-1}X_j$ where $Z \sim \mathrm{N}(0,1)$ independent of $\mathbf{X}$. Note that for the two heteroscedastic distributions, we have $\mathrm{Cov}(\mathbf{X},\epsilon) = \mathrm{E}(\epsilon\mathbf{X}) = \mathbf{0}$, although ϵ is not independent with $\mathbf{X}$. For the first heteroscedastic case the conditional variance only depends on a low dimensional covariates and the conditional variance for the second heteroscedastic case depends on the the entire vector of covariates. The true coefficients $\boldsymbol{\beta}^0$ satisfies $\beta_1^0 = 0, 0.1, 0.2, 0.3, 0.4, 0.5$ (0 for the size and others for the power analysis), $\beta_4^0 = 1.5, \beta_7^0 = 2$ and all others are 0. Our goal is to test

$$H_0 : \beta_1^0 = 0, \text{ v.s. } H_1 : \beta_1^0 \neq 0.$$

The number of simulations is 500.

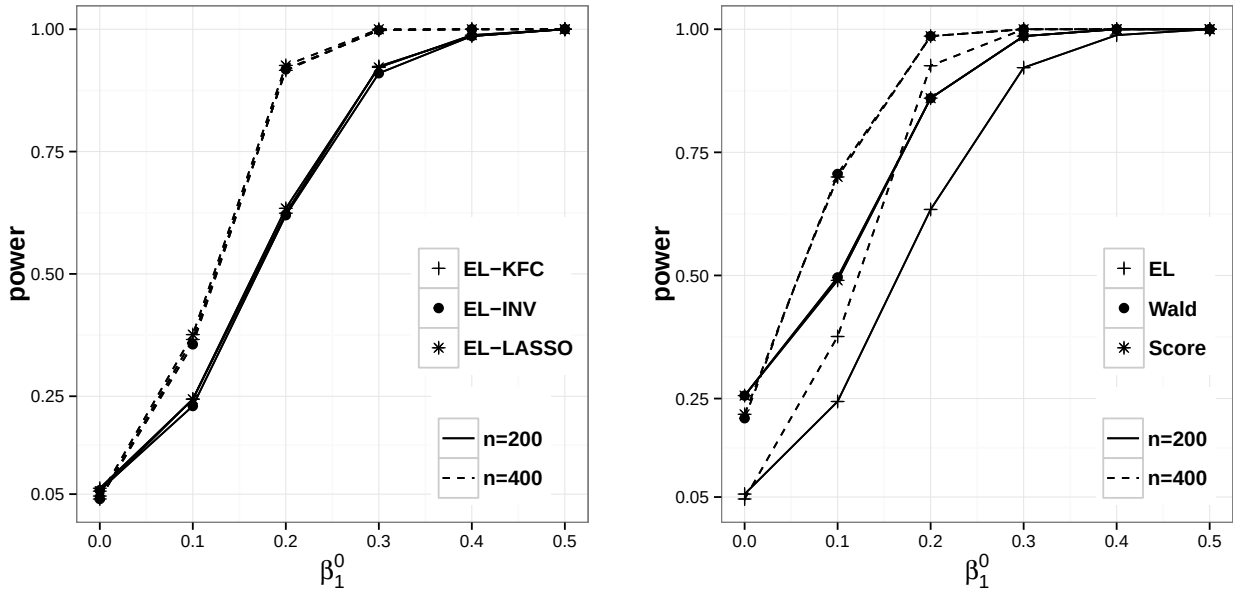
For the initial estimators such as $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}_1$ and $\hat{\mathbf{w}}_1$, we just use the scaled Lasso [SZ12], which has the advantage of being tuning insensitive. "EL-KFC" corresponds to the KFC Projection example, "EL-INV" corresponds to the Inverse Projection example, and "EL-LASSO" corresponds to Lasso Projection example. And "Wald" corresponds to the Wald type test as proposed in [ZZ14] and [vdGBR13], while "Score" corresponds to the Score type test as proposed in [NL14] with Lasso estimation for $\hat{\mathbf{w}}_1$.

And for the "EL-KFC", in order to stabilize the KFC set selection, we used the stability selection procedure through sub-sampling proposed by [MB10] and [SS13]. According to [SS13], we split the data into half for 10 times and select the final KFC set with variables shown at least 50% of those 20 KFC sets.

For illustration, we only show some of the cases here. In Table 4.1 with Toeplitz matrix with $\rho = 0.2$ as the covariance matrix for the predictors and standard normal error, we can see that all of the procedures has reasonably well controlled type I error around α level at

5%. And for the empirical likelihood based approach with different estimating equations, they have pretty much similar power performance. An interesting comparison among the holy trinity, i.e. Wald type test, Score test and the likelihood ratio test, which correspond to the last three sections in Table 4.1, shows that the likelihood ratio test has overall better power performance than the other two, especially in the low sample size situation.

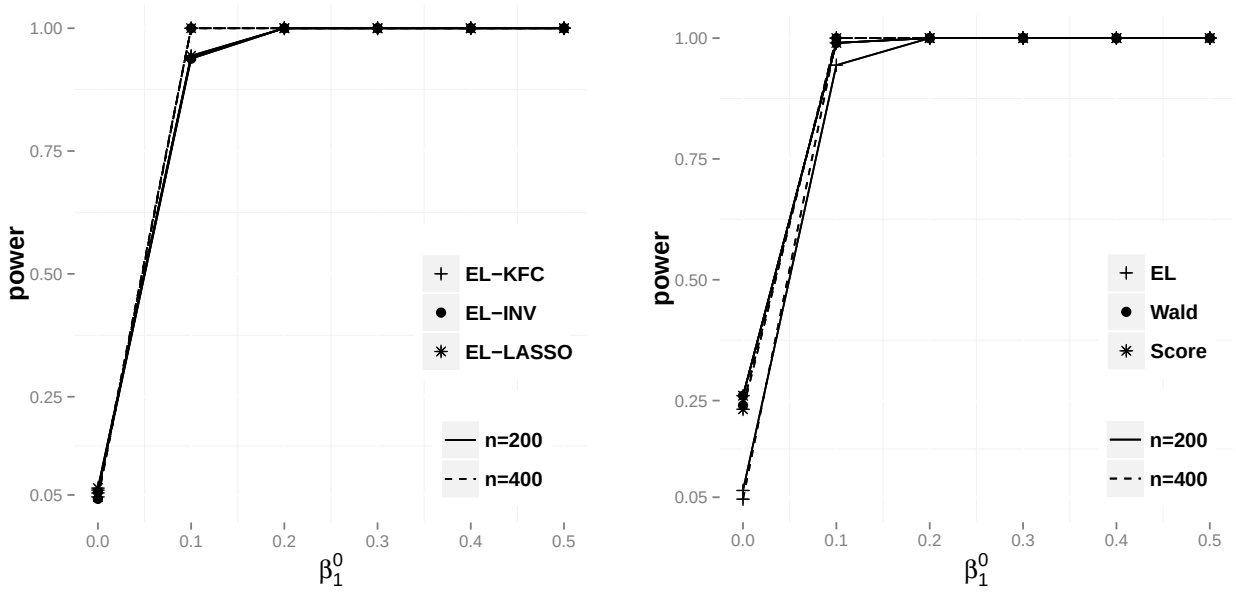
The most exciting part is about the heteroscedasticity. In Table 4.2, we simulate the predictors with the Toeplitz covariance matrix with $\rho = 0.2$ and the heteroscedastic noise $0.7X_1\mathrm{N}(0,1)$. Under this case, we could see clearly that all of the empirical likelihood based inference procedures, which corresponds to the first four sections in Table 4.2, are valid, i.e. they have reasonably well controlled type I error. The same results are also demonstrated in Figure 4.1a for $p = 100$. But for the Wald type test and Score test, their type I errors are largely inflated, which indicates that these two procedures are invalid. We can clearly see the patterns in Figure 4.1b for $p = 100$. For the other heteroscedastic noise with conditional error variance depending on high dimensional covariates, that is $\frac{1}{p-1}X_1\sum_{j=2}^{p}X_{j-1}X_j\mathrm{N}(0,1)$, we can observe similar performances in Table 4.3, as well as in Figure 4.2 for $p = 500$. This shows the advantage of the empirical likelihood based inference procedures.



(a) Empirical Likelihood Based

(b) Holy Trinity

Figure 4.1: **Empirical Size and Power Comparison among Empirical Likelihood based approaches and among Holy Trinity and $p = 100$.** (a) "EL-KFC" represents EL approach with KFC projection, "EL-INV" represents EL approach with inverse projection and "EL-LASSO" represents EL approach with Lasso projection; (b) "Wald" represents Wald type test, "Score" represents Score test and "EL" represents likelihood ratio test.



(a) Empirical Likelihood Based

(b) Holy Trinity

Figure 4.2: **Empirical Size and Power Comparison among Empirical Likelihood based approaches and among Holy Trinity with Heteroscedastic Noise** $\frac{1}{p-1}X_1\sum_{j=2}^{p}X_{j-1}X_j\mathbf{N}(0,1)$ **and** $p = 500$**.** (a) "EL-KFC" represents EL approach with KFC projection, "EL-INV" represents EL approach with inverse projection and "EL-LASSO" represents EL approach with Lasso projection; (b) "Wald" represents Wald type test, "Score" represents Score test and "EL" represents likelihood ratio test.

Table 4.1: **Power comparison.** Covariate: Toeplitz matrix with $\rho = 0.2$; Error: $N(0, 1)$.

			β_1^0					
Method	p	n	0	0.1	0.2	0.3	0.4	0.5
EL-KFC	100	200	0.054	0.304	0.760	0.984	1	1
		400	0.052	0.482	0.964	1.000	1	1
	200	200	0.052	0.294	0.762	0.976	1	1
		400	0.044	0.460	0.980	1.000	1	1
	500	200	0.064	0.292	0.760	0.972	1	1
		400	0.040	0.488	0.972	1.000	1	1
EL-INV	100	200	0.040	0.296	0.748	0.984	1	1
		400	0.054	0.470	0.962	1.000	1	1
	200	200	0.044	0.290	0.774	0.976	1	1
		400	0.038	0.458	0.980	1.000	1	1
	500	200	0.048	0.276	0.784	0.978	1	1
		400	0.034	0.490	0.972	1.000	1	1
EL-LASSO	100	200	0.052	0.312	0.770	0.990	1	1
		400	0.054	0.490	0.970	1.000	1	1
	200	200	0.048	0.308	0.786	0.982	1	1
		400	0.038	0.462	0.978	1.000	1	1
	500	200	0.056	0.300	0.788	0.980	1	1
		400	0.042	0.512	0.976	1.000	1	1
Wald	100	200	0.048	0.266	0.748	0.964	1.000	1
		400	0.048	0.502	0.970	1.000	1.000	1
	200	200	0.064	0.270	0.742	0.972	1.000	1
		400	0.038	0.486	0.978	1.000	1.000	1
	500	200	0.052	0.284	0.794	0.968	0.998	1
		400	0.040	0.486	0.978	1.000	1.000	1
Score	100	200	0.050	0.264	0.746	0.962	1.000	1
		400	0.052	0.480	0.966	1.000	1.000	1
	200	200	0.062	0.268	0.740	0.970	1.000	1
		400	0.040	0.474	0.978	1.000	1.000	1
	500	200	0.062	0.272	0.794	0.970	0.998	1
		400	0.038	0.498	0.976	1.000	1.000	1

Table 4.2: **Power comparison.** Covariate: Toeplitz matrix with $\rho = 0.2$; Error: $0.7X_1\mathrm{N}(0,1)$.

			β_1^0					
Method	p	n	0	0.1	0.2	0.3	0.4	0.5
EL-KFC	100	200	0.062	0.244	0.624	0.924	0.986	1.000
		400	0.040	0.366	0.916	0.998	1.000	1.000
	200	200	0.070	0.230	0.652	0.920	0.990	1.000
		400	0.076	0.350	0.890	0.990	1.000	1.000
	500	200	0.060	0.254	0.636	0.900	0.986	0.996
		400	0.058	0.402	0.902	0.992	1.000	1.000
EL-INV	100	200	0.058	0.230	0.620	0.910	0.986	1.000
		400	0.040	0.356	0.918	0.998	1.000	1.000
	200	200	0.058	0.222	0.652	0.910	0.988	1.000
		400	0.066	0.342	0.880	0.990	1.000	1.000
	500	200	0.060	0.236	0.624	0.898	0.980	0.996
		400	0.050	0.402	0.902	0.992	1.000	1.000
EL-LASSO	100	200	0.056	0.244	0.634	0.922	0.988	1.000
		400	0.046	0.376	0.926	1.000	1.000	1.000
	200	200	0.062	0.232	0.668	0.926	0.990	1.000
		400	0.072	0.356	0.890	0.988	1.000	1.000
	500	200	0.068	0.250	0.640	0.912	0.986	0.996
		400	0.052	0.412	0.902	0.992	1.000	1.000
Wald	100	200	0.256	0.496	0.860	0.986	1	1
		400	0.210	0.706	0.986	1.000	1	1
	200	200	0.234	0.464	0.848	0.980	1	1
		400	0.236	0.680	0.968	1.000	1	1
	500	200	0.208	0.516	0.874	0.978	1	1
		400	0.234	0.736	0.986	1.000	1	1
Score	100	200	0.256	0.490	0.860	0.986	1	1
		400	0.218	0.700	0.986	1.000	1	1
	200	200	0.234	0.470	0.846	0.980	1	1
		400	0.234	0.672	0.968	1.000	1	1
	500	200	0.204	0.518	0.870	0.978	1	1
		400	0.230	0.728	0.984	1.000	1	1

Table 4.3: **Power comparison.** Covariate: Toeplitz matrix with $\rho = 0.2$; Error: $\frac{1}{p-1}X_1 \sum_{j=2}^{p} X_{j-1}X_j \mathrm{N}(0,1)$.

			β_1^0					
Method	p	n	0	0.1	0.2	0.3	0.4	0.5
EL-KFC	100	200	0.066	0.886	0.998	1	1	1
		400	0.048	0.988	1.000	1	1	1
	200	200	0.076	0.932	1.000	1	1	1
		400	0.068	0.988	1.000	1	1	1
	500	200	0.060	0.942	1.000	1	1	1
		400	0.054	1.000	1.000	1	1	1
EL-INV	100	200	0.062	0.872	0.998	1	1	1
		400	0.038	0.988	1.000	1	1	1
	200	200	0.074	0.936	1.000	1	1	1
		400	0.064	0.988	1.000	1	1	1
	500	200	0.056	0.938	1.000	1	1	1
		400	0.042	1.000	1.000	1	1	1
EL-LASSO	100	200	0.066	0.876	0.998	1	1	1
		400	0.046	0.988	1.000	1	1	1
	200	200	0.078	0.934	1.000	1	1	1
		400	0.064	0.988	1.000	1	1	1
	500	200	0.064	0.944	1.000	1	1	1
		400	0.046	1.000	1.000	1	1	1
Wald	100	200	0.222	0.982	1	1	1	1
		400	0.214	1.000	1	1	1	1
	200	200	0.244	0.990	1	1	1	1
		400	0.214	0.998	1	1	1	1
	500	200	0.260	0.990	1	1	1	1
		400	0.240	1.000	1	1	1	1
Score	100	200	0.226	0.984	1	1	1	1
		400	0.208	1.000	1	1	1	1
	200	200	0.236	0.990	1	1	1	1
		400	0.206	0.998	1	1	1	1
	500	200	0.260	0.990	1	1	1	1
		400	0.232	1.000	1	1	1	1

4.6 Real Data Analysis

Microarray expression experiments and array-based comparative genomic hybridization (array CGH) experiments have been conducted for more than 170 primary breast tumor specimens in a few recent breast cancer cohort studies, collected at multiple cancer centers [FFS10].

We used a total of 172 tumor samples with both cDNA expression microarray and CGH array data. In our study, we used the copy number alteration intervals (CNAIs), which are defined as basic CNA units (genome regions) in which all genes tend to be amplified or deleted simultaneously in a sample. For each CNAI in each sample, the mean value of the estimated copy numbers of the genes falling into this CNAI was calculated. This resulted in a 172 (samples) by 384 (CNAIs) numeric matrix. After global normalization for each expression array, we focused on a set of 654 breast cancer related genes, which was derived based on seven published breast cancer gene lists. This resulted in a 172 (samples) by 654 (genes) numeric matrix. See model details about the data processing in [PZB$^+$10].

Our study tends to reveal the subtle and complicated regulatory relationships among DNA copy numbers and RNA transcript levels. The dependence of RNA levels on DNA copy numbers can be modeled through a straightforward multivariate linear regression model with the RNA levels as responses and the DNA copy numbers as predictors. While multivariate linear regression is well studied in statistical literature, the current problem bears new challenges due to high-dimensionality in terms of both predictors and responses. We will adopt some dimension reduction procedures for the RNA expressions followed by significant association detect using the proposed methods.

4.6.1 WGCNA of correlated genes

In order to deal with the correlation patterns among genes across microarray samples, we adopted the Weighted Gene Co-expression Network Analysis (WGCNA) [LH08] which can be used for finding clusters (modules) of highly correlated genes.

By using of WGCNA with `minModuleSize=10`, we identified 5 modules labeled 1 through 5 in order of descending size as listed in Table 4.4. The label 0 is reserved for genes outside of all modules. And we can see that in Figure 4.3a there are pretty clear 5 modules clustered.

Table 4.4: **Module Sizes.**

module	0	1	2	3	4	5
size	204	316	53	33	25	23

For summarizing such cluster, we use the module eigengene by conducting PCA for each of the five modules to select the first principal component as our response to do the association analysis. But since we have missing values in the expression data matrix, we impute the missing values by using `impute.knn` [HTS+99]. And we choose `prcomp` in `R` to perform PCA.

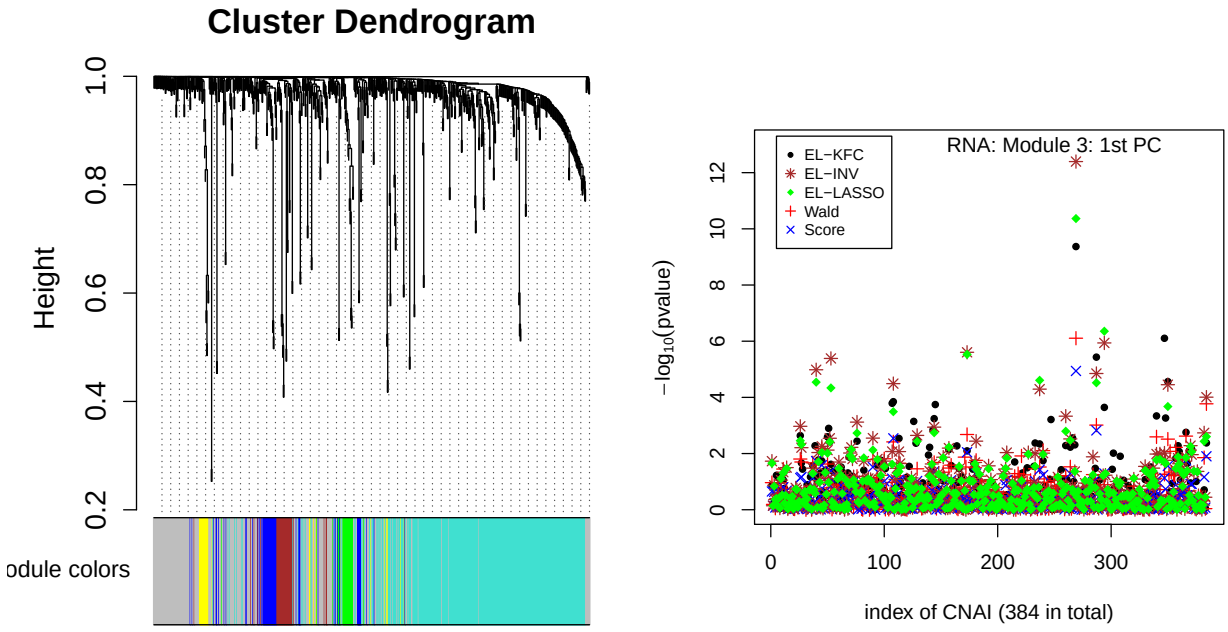
4.6.2 Significance Test

After doing WGCNA, imputation and PCA, we have first principal component for each of 5 modules. We regressed each module eigengene onto the predictors (CNAIs, totally we have 384) separately to conduct the single coefficient significance study.

For illustration, we just demonstrate the results for Module 3 and see others in the appendix. In Figure 4.3b, we could see that although different methods have different power

performance, the significance spots are very much consistent all over the methods. And the Score method is kind of powerless in this data analysis among all 6 different methods.

For each inference procedure testing for all covariates, we can get a sequence of p-values $\{p_j\}_{j=1}^{p}$. With the ordered p-values, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(j)} \leq \cdots \leq p_{(p)}$, we adopt the Benjamini-Hochberg (BH) algorithm: for a fixed value $\alpha = 0.01\%$, let $j_{\max}$ be the largest index for which $p_{(j)} \leq \frac{j}{p}\alpha$, and reject $H_{0(j)}$, the null hypothesis corresponding to $p_{(j)}$, if $j \leq j_{\max}$, accepting $H_{0(j)}$ otherwise. Take Module 3 as an example, we found that all of the empirical likelihood based approaches detected one consistent signal which is the 269-th CNAI, on Chromosome 15 with Cytoband "15q11.2-15q11.2".



(a) Clustering

(b) Manhattan Plot

Figure 4.3: **Breast Cancer Cohort Studies.** (a) Clustering dendrogram of genes, with dissimilarity based on topological overlap, together with assigned module colors. (b) Manhattan plot for Module 3.

4.6.3 Presence of Heteroscedasticity

We test for the presence of heteroscedasticity in this data set for each of the 654 genes using the Goldfeld-Quandt test [GQ65]. The Goldfeld-Quandt test is one of the most widely used test for heteroscedasticity. It compares the variances of two submodels divided by a specified breakpoint and rejects if the variances differ. The Goldfeld-Quandt test is not directly applicable when $p > n$. To reduce the dimensions, we apply the $\sqrt{\text{Lasso}}$ to select CNAIs that are predictive of gene expression levels and CNAIs that are explanatory of variability. These variables are then applied in the Goldfeld-Quandt test to specify predictors on the response. Since the $\sqrt{\text{Lasso}}$ is not that sensitive to the selection of the tuning parameter and we are also durable to select more variables, we just set the tuning parameter to be $\sqrt{\log p/n}$.

We found that 19 out of 654 genes demonstrate heteroscedasticity at the significance level 0.05/654. The presence of heteroscedasticity for these genes suggests the need to use our method for identifying the CNAIs that are associated with gene expression. As further evidence for the existence of heteroscedasticity, we apply the "wandering schematic plot" [Tuk77]. This slices the predicted value into bins and uses m-letter summaries (generalizations of boxplots) to show the location, spread, and shape of the residuals for each bin. The m-letter statistics are further smoothed in order to emphasize overall patterns rather than chance deviations. Figure 4.4 presents the "wandering schematic plots" for genes PDK3 (Chr 23), TPST2 (Chr 22), ELF3 (Chr 1) and SNRPE (Chr 22), which are the top 4 genes for the heteroscedasticity.

4.6.4 Results for Top 4 Genes with Heteroscedasticity

We apply our Empirical Likelihood based approach to the four genes discussed in the previous section and demonstrated in Figure 4.5 and compare its performances with those of Wald type test and Score type test. For example, we use gene TPST2 on Chr 22 as shown in Figure 4.5b for demonstration. For each inference procedure testing for all covariates, we can get a sequence of p-values $\{p_j\}_{j=1}^{p}$. With the ordered p-values, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(j)} \leq \cdots \leq p_{(p)}$, we adopt the Benjamini-Hochberg (BH) algorithm to make the decision. As a result, we found that only EL-INV and EL-LASSO can detect signals and all of the other procedures found nothing significant. Moreover EL-INV and EL-LASSO found two consistent signals at the 305-th CNAI and 307-th CNAI, both of which are on Chromosome 17 with Cytoband "17q12-17q12".

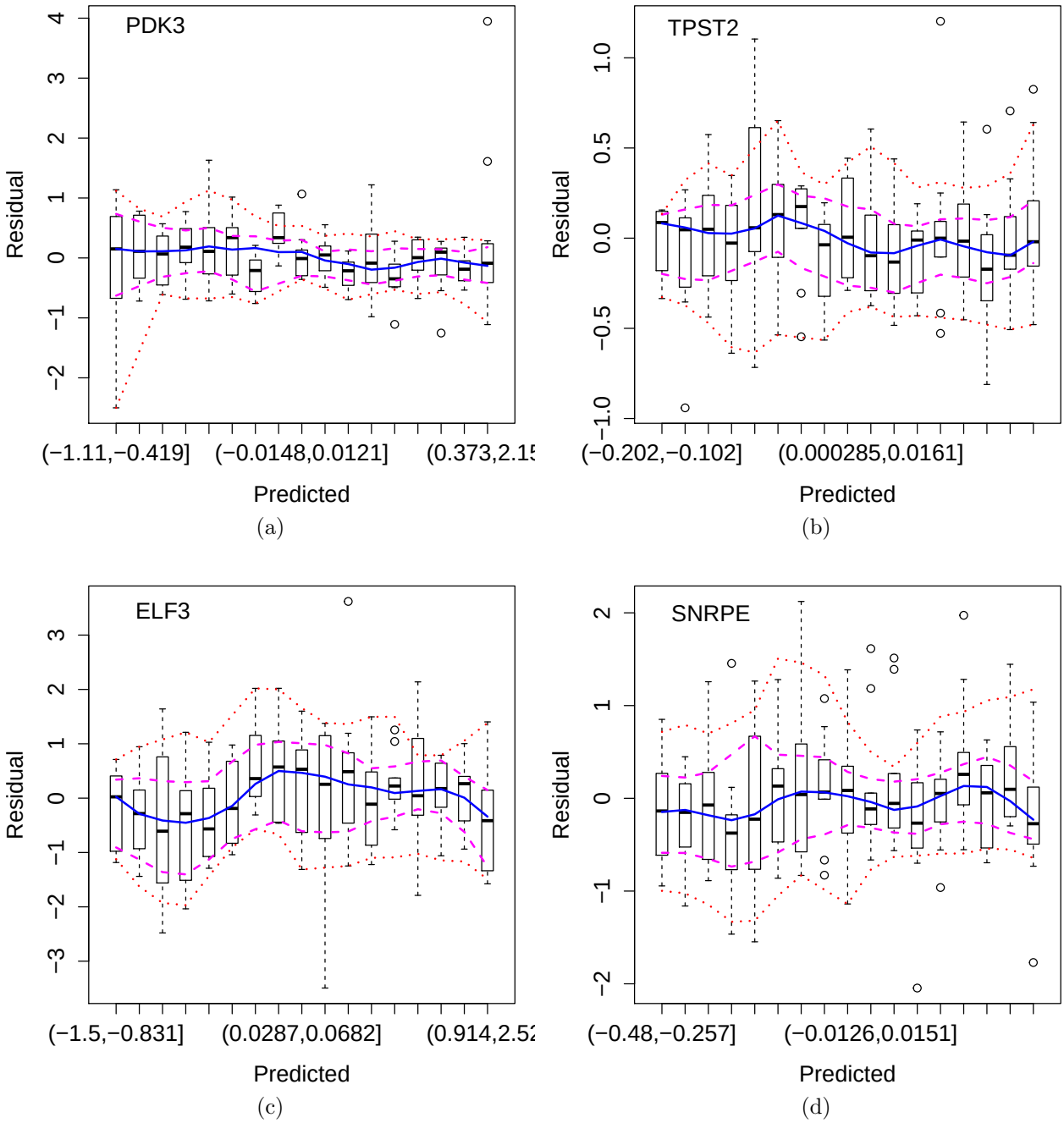
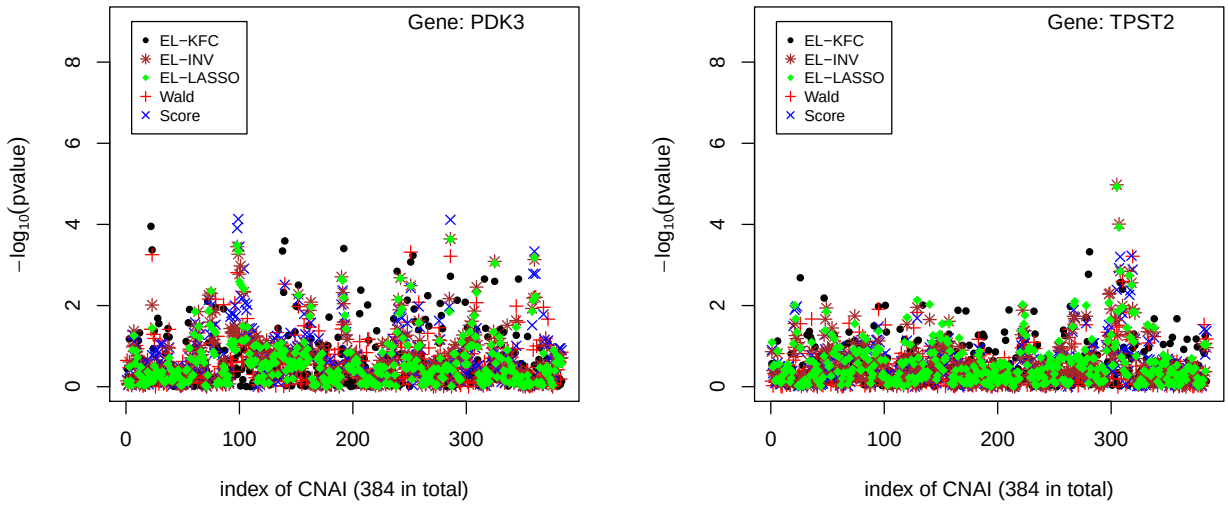
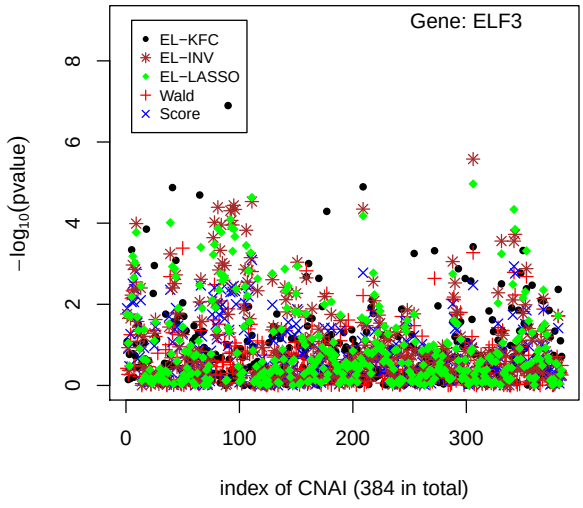


Figure 4.4: **Wondering Schematic Plot for Top 4 Genes with Heteroscedasticity.**

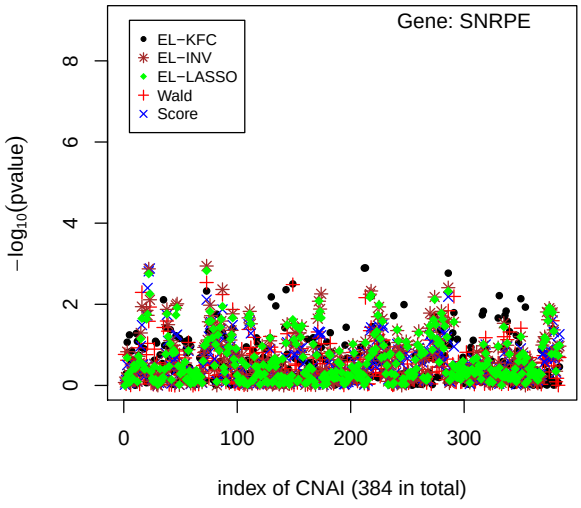


(a)

(b)



(c)



(d)

Figure 4.5: **Manhattan Plot for Top 4 Genes with Heteroscedasticity.**

4.7 Technical Details

4.7.1 Assumptions for Theoretical Examples

Assumption 1. *(1) Assume the initial estimator $\hat{\boldsymbol{\beta}}$ satisfying $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$.*

(2) Suppose the initial estimators $\hat{\mathbf{w}}_j$ satisfy $\max_{1\le j\le p}\|\hat{\mathbf{w}}_j - \mathbf{w}_j^0\|_1 = O_p(a_n)$, where $a_n = o(1/\sqrt{\log p})$.

(3) The prediction errors satisfy $\|\mathbb{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2/n = O_p(s\log p/n)$ and $\max_{1\le j\le p}\|\mathbb{X}_{\setminus j}(\hat{\mathbf{w}}_j - \mathbf{w}_j^0)\|_2^2/n = O_p(b_n)$, where $\mathbb{X}_{\setminus j}$ is the design matrix $\mathbb{X}$ with the j-th column deleted and $b_n = o(1/\sqrt{n})$.

(4) $\mathbf{X}_i$ and ϵ_i are all sub-Gaussian.

(5) $s\log p/\sqrt{n} = o(1)$.

Remark 6. *1. With (4) that $\mathbf{X}_i$ and ϵ_i are all sub-Gaussian, we have $X_{ik}\epsilon_i$ sub-exponential with $E(\epsilon_i X_{ik}) = 0$. By Bernstein inequality [Ver10] and union bound inequality, we have*

$$P(\big\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\epsilon_i\big\|_\infty \ge t) \le C_1 p\exp(-C\min(t^2/C_2, t/C_3)n).$$

By taking $t = C'\sqrt{\frac{\log p}{n}}$ for some positive constant C' such that $CC'^2 > C_2$, we have

$$\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\epsilon_i\|_\infty = O_p(\sqrt{\frac{\log p}{n}}). \tag{4.7.21}$$

2. For $\eta_{ij} = X_{ij} - E(X_{ij}|\mathbf{X}_{i,\setminus j})$, we have η_{ij} sub-gaussian since $\mathbf{X}_i$ is sub-gaussian. And for any $k \ne j$, we have $E(X_{ik}\eta_{ij}) = E\{X_{ik}[X_{ij} - E(X_{ij}|\mathbf{X}_{i,\setminus j})]\} = E\{X_{ik}X_{ij} -$

$E[X_{ik}X_{ij}|\mathbf{X}_{i,\backslash j}]\} = 0$. *Similarly, we have for any* $t > 0$ *and* $1 \leq j \neq k \leq p$,

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_{ik}\eta_{ij}| \geq t) \leq C_1 p \exp(-C\min(t^2/C_2, t/C_3)n),$$

which leads to

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\mathbf{X}_{i,\backslash j}\Big\|_{\infty} = O_p(\sqrt{\frac{\log p}{n}}). \tag{4.7.22}$$

3. *For the properties of the initial estimators in (1), (2) and (3) under the heteroscedasitic noise case, we can use the* $\sqrt{Lasso}$ *estimator as in [BCW14]. According to Theorem 7 in [BCW14], we have that the* $\sqrt{Lasso}$ *estimators under certain conditions have these properties satisfied.*

Assumption 2. *(1) Assume the same assumption as Lasso projection case for the initial estimator* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$.

(2) Assume similar assumption as Lasso projection case for the initial estimators $\hat{\boldsymbol{\gamma}}_j$, *i.e.* $\max_{1\leq j\leq p}\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0\|_1 = O_p(a_n)$, *where* $a_n = o(1/\sqrt{\log p})$.

(3) Assume similar assumption as Lasso projection case for the prediction errors, i.e. $\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n = O_p(s\log p/n)$ *and* $\max_{1\leq j\leq p}\|(\mathbb{Y}, \mathbb{X}_{\backslash j})(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0)\|_2^2/n = O_p(b_n)$ *and* $b_n = o(1/\sqrt{n})$.

(4) $(\mathbf{X}_i^{\intercal}, \epsilon_i)^{\intercal}$ *is sub-Gaussian.*

(5) $s\log p/\sqrt{n} = o(1)$.

Remark 7. *For the condition (2) above, if we assume* $a = \max_{1\leq j\leq p} s_j$ *with* $s_j = \|\boldsymbol{\gamma}_j^0\|_0$ *and then the* $\sqrt{Lasso}$ *estimators for* $\boldsymbol{\gamma}_j^0$ *satisfy this condition with* $a_n = a\sqrt{\log p/n}$. *For the*

condition (3) above, since we assume that $(\mathbf{X}_i^\intercal, \epsilon_i)^\intercal$ *is sub-Gaussian (which makes* $\boldsymbol{\beta}^{0\intercal}\mathbf{X}_i$ *also sub-Gaussian), then due to* $Cov(\boldsymbol{\beta}^{0\intercal}\mathbf{X}_i, \epsilon_i) = E(\epsilon_i \boldsymbol{\beta}^{0\intercal}\mathbf{X}_i) = 0$*, we have* $\epsilon_i \boldsymbol{\beta}^{0\intercal}\mathbf{X}_i$ *sub-exponential and by the Bernstein inequality, we have for any* $t > 0$*,*

$$P(\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\intercal\boldsymbol{\beta}^0\epsilon_i\Big| \geq t) \leq 2\exp\{-C_1 n \min(t^2/C_2^2, t/C_2)\}.$$

This also leads to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\intercal\boldsymbol{\beta}^0\epsilon_i = O_p(\sqrt{\log p/n}), \tag{4.7.23}$$

as long as $\log p/n \to 0$*. And with the same argument, we have*

$$\frac{1}{n}\sum_{i=1}^{n}X_{ik}\eta_{ij,y} = O_p(\sqrt{\log p/n}), \tag{4.7.24}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i, \mathbf{X}_{i,\setminus j}^\intercal)\boldsymbol{\gamma}_j^0\eta_{ij,y} = O_p(\sqrt{\log p/n}). \tag{4.7.25}$$

Assumption 3. *(1) For the eigenvalues of* $\boldsymbol{\Sigma}$*, there exist some constants* $\lambda_{\min}$ *and* $\lambda_{\max}$ *such that*

$$0 < \lambda_{\min} < \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) < \lambda_{\max} < \infty.$$

(2) Assume $\mathbf{X}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ *and* ϵ_i *to be sub-Gaussian.*

(3) Assume the same as the Lasso projection for the initial estimator $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$*.*

(4) $m^3 \log p/n = o(1)$*,* $s\sqrt{\frac{(\log p)^2 m^3}{n}} = o(1)$*,* $s\sqrt{\frac{(\log p)^3 m^2}{n^2}} = o(1)$*.*

(5) Assume $s\sqrt{\log p}\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{k\in\mathcal{S}^*}\left|\sigma_{jk} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}k}\right| = o(1)$ *to control the partial correlation between the target covariate* X_{ij} *and* $\mathbf{X}_{i\mathcal{S}^*}$*.*

4.7.2 Proof of Theorems

Proof of Theorem 5. As in [Owe01], by (C0), with probability tending to 1, $-2\log \mathrm{EL}_n(\beta_j^0) = 2\sum_{i=1}^n \log(1+\lambda m_{ni})$ where λ satisfies

$$\sum_{i=1}^n \frac{m_{ni}}{1+\lambda m_{ni}} = 0. \tag{4.7.26}$$

The next step is to bound the magnitude of λ. Let $\lambda = |\lambda|u$ where $u = \mathrm{sign}(\lambda) \in \{-1, 1\}$. Now by $\sum_{i=1}^n \frac{m_{ni}}{1+\lambda m_{ni}} = 0$, we have

$$0 = \sum_{i=1}^n \frac{um_{ni}}{1+\lambda m_{ni}} = \sum_{i=1}^n um_{ni}\big\{1 - \frac{\lambda m_{ni}}{1+\lambda m_{ni}}\big\},$$

which implies

$$\sum_{i=1}^n um_{ni} = \sum_{i=1}^n \frac{u\lambda m_{ni}^2}{1+\lambda m_{ni}} = \sum_{i=1}^n \frac{|\lambda| m_{ni}^2}{1+\lambda m_{ni}} \geq |\lambda| \sum_{i=1}^n \frac{m_{ni}^2}{1+|\lambda| \max_{1\leq i\leq n} |m_{ni}|}.$$

Thus we have

$$u\frac{1}{n}\sum_{i=1}^n m_{ni} \geq \frac{|\lambda|}{1+|\lambda| \max_{1\leq i\leq n} |m_{ni}|} \frac{1}{n}\sum_{i=1}^n m_{ni}^2.$$

which implies

$$|\lambda|\big\{\frac{1}{n}\sum_{i=1}^n m_{ni}^2 - (\max_{1\leq i\leq n} |m_{ni}|)u\frac{1}{n}\sum_{i=1}^n m_{ni}\big\} \leq u\frac{1}{n}\sum_{i=1}^n m_{ni}. \tag{4.7.27}$$

From (C1), by Lemma 3 in [Owe90], we have $\max_{1\le i\le n}|W_{ni}| = o_p(n^{1/2})$, and together with (C2), we have

$$\max_{1\le i\le n}|m_{ni}| = o_p(n^{1/2}). \tag{4.7.28}$$

And since for any $\epsilon > 0$,

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathrm{E}\big\{W_{ni}^2\mathbb{1}(|W_{ni}| > \epsilon\sqrt{n}\sigma_n)\big\} = \sigma_n^{-2}\mathrm{E}\big\{W_{n1}^2\mathbb{1}(|W_{n1}| > \epsilon\sqrt{n}\sigma_n)\big\},$$

where obviously $W_{n1}^2\mathbb{1}(|W_{n1}| > \epsilon\sqrt{n}\sigma_n) \xrightarrow{p} 0$ due to $\mathrm{P}(|W_{n1}| > \epsilon\sqrt{n}\sigma_n) \to 0$, we have by Dominated Convergence Theorem,

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathrm{E}\big\{W_{ni}^2\mathbb{1}(|W_{ni}| > \epsilon\sqrt{n}\sigma_n)\big\} \to 0.$$

Thus by Lindeberg-Feller Central Limit Theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ni} \xrightarrow{d} \mathrm{N}(0,\sigma_w^2). \tag{4.7.29}$$

By (4.7.29) and together with (C2), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m_{ni} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ni} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}R_{ni} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ni} + o_p(1) \xrightarrow{d} \mathrm{N}(0,\sigma_w^2). \tag{4.7.30}$$

And by (C1) and (C2) we have

$$\frac{1}{n}\sum_{i=1}^{n}m_{ni}^2 = \frac{1}{n}\sum_{i=1}^{n}W_{ni}^2 + \frac{1}{n}\sum_{i=1}^{n}R_{ni}^2 + 2\frac{1}{n}\sum_{i=1}^{n}W_{ni}R_{ni} = \frac{1}{n}\sum_{i=1}^{n}W_{ni}^2 + o_p(1) \to \sigma_w^2. \tag{4.7.31}$$

Actually the above follows from checking the WLLN for triangular arrays. First of all $\sum_{i=1}^n \mathrm{P}(W_{ni}^2 > n) = n\mathrm{P}(W_{n1}^2 > n) \le \mathrm{E}\{W_{n1}^2 \mathbb{1}(W_{n1}^2 > n)\} \to 0$; and

$$\begin{aligned} n^{-2}\sum_{i=1}^n \mathrm{E}\{W_{ni}^4 \mathbb{1}(W_{ni}^2 \le n)\} &= n^{-1}\mathrm{E}\{W_{n1}^4 \mathbb{1}(W_{n1}^2 \le n)\} \\ &= n^{-1}\int_0^n 2y\mathrm{P}(W_{n1}^2 > y)dy \to 0 \end{aligned}$$

since $y\mathrm{P}(W_{n1}^2 > y) \le \mathrm{E}(W_{n1}^2 \mathbb{1}(W_{n1}^2 > y)) \to 0$ as $y \to \infty$.

Thus by (4.7.27), (4.7.28), (4.7.30) and (4.7.31), we have

$$|\lambda|(\frac{1}{n}\sum_{i=1}^n m_{ni}^2 + o_p(1)) = O_p(n^{-1/2})$$

and hence

$$|\lambda| = O_p(n^{-1/2}). \tag{4.7.32}$$

Then it follows from (4.7.28), we have $\max_{1\le i\le n}\left|\frac{\lambda m_{ni}}{1+\lambda m_{ni}}\right| = o_p(1)$. Therefore, from (4.7.26), we have

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^n \frac{\lambda m_{ni}}{1+\lambda m_{ni}} = \frac{1}{n}\sum_{i=1}^n \lambda m_{ni}\Big\{1 - \lambda m_{ni} + \frac{[\lambda m_{ni}]^2}{1+\lambda m_{ni}}\Big\} \\ &= \frac{1}{n}\sum_{i=1}^n \lambda m_{ni} - \frac{[1+o_p(1)]}{n}\sum_{i=1}^n [\lambda m_{ni}]^2, \end{aligned}$$

which leads to

$$\frac{1}{n}\sum_{i=1}^n \lambda m_{ni} = \frac{[1+o_p(1)]}{n}\sum_{i=1}^n [\lambda m_{ni}]^2. \tag{4.7.33}$$

Again by using (4.7.26) and together with (4.7.30), we have

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{m_{ni}}{1+\lambda m_{ni}} = \frac{1}{n}\sum_{i=1}^{n} m_{ni}\Big\{1-\lambda m_{ni}+\frac{[\lambda m_{ni}]^2}{1+\lambda m_{ni}}\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} m_{ni} - \frac{\lambda}{n}\sum_{i=1}^{n} m_{ni}^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{m_{ni}[\lambda m_{ni}]^2}{1+\lambda m_{ni}} \\
&= \frac{1}{n}\sum_{i=1}^{n} m_{ni} - \frac{\lambda}{n}\sum_{i=1}^{n} m_{ni}^2 + O_p\Big\{\max_{1\le i\le n}\big|\frac{m_{ni}}{1+\lambda m_{ni}}\big|\frac{1}{n}\sum_{i=1}^{n}[\lambda m_{ni}]^2\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} m_{ni} - \frac{\lambda}{n}\sum_{i=1}^{n} m_{ni}^2 + o_p\Big\{n^{1/2}\lambda^2\frac{1}{n}\sum_{i=1}^{n} m_{ni}^2\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} m_{ni} - \frac{\lambda}{n}\sum_{i=1}^{n} m_{ni}^2 + o_p(n^{-1/2}),
\end{aligned}$$

which leads to

$$\lambda = \Big\{\frac{1}{n}\sum_{i=1}^{n} m_{ni}^2\Big\}^{-1}\frac{1}{n}\sum_{i=1}^{n} m_{ni} + o_p(n^{-1/2}). \tag{4.7.34}$$

Finally, by Taylor expansion together with (4.7.30), (4.7.31), (4.7.33) and (4.7.34), we have

$$\begin{aligned}
-2\log \mathrm{EL}_n(\beta_j^0) &= 2\sum_{i=1}^{n}\log(1+\lambda m_{ni}) \\
&= 2\sum_{i=1}^{n}\lambda m_{ni} - [1+o_p(1)]\sum_{i=1}^{n}[\lambda m_{ni}]^2 \\
&= [1+o_p(1)]\sum_{i=1}^{n}[\lambda m_{ni}]^2 = [1+o_p(1)]\lambda^2\sum_{i=1}^{n} m_{ni}^2 \\
&= [1+o_p(1)]\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_{ni}\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} m_{ni}^2\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_{ni}\Big) + o_p(1) \\
&\xrightarrow{d} \chi_1^2, \text{ as } n\to\infty.
\end{aligned}$$

This completes the proof of the theorem. □

Proof of Theorem 6. We only need to control the term R_{ni}, which will be controlled one by one.

By (3) in Assumption 1, we have (4.7.21) and (4.7.22), which leads to

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n} R_{ni,1}\Big| &= \Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^\intercal\boldsymbol{\beta}^0)(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal\mathbf{X}_{i,\setminus j}\Big| \\ &= \Big|(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,\setminus j}\epsilon_i\Big| \le \|\mathbf{w}_j^0 - \hat{\mathbf{w}}_j\|_1\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,\setminus j}\epsilon_i\|_\infty \\ &= O_p(a_n)O_p(\sqrt{\frac{\log p}{n}}) = O_p(a_n\sqrt{\frac{\log p}{n}}).\end{aligned}$$

In order to have $\Big|\frac{1}{n}\sum_{i=1}^{n} R_{ni,1}\Big| = o_p(n^{-1/2})$ we need to have $a_n = o(1/\sqrt{\log p})$, which is true according to (2) in Assumption 1.

For $R_{ni,2}$, we have

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n} R_{ni,2}\Big| = \Big|\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \hat{\mathbf{w}}_j^\intercal\mathbf{X}_{i,\setminus j})\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\Big| \\ =&\Big|\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j}) + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal\mathbf{X}_{i,\setminus j}\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\Big| \\ \le&\Big|\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal\mathbf{X}_{i,\setminus j}\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\Big| \\ \le&\Big\|\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\mathbf{X}_{i,\setminus j}^\intercal\Big\|_\infty\Big\|\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\Big\|_1 \\ &+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\big\{(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^\intercal\mathbf{X}_{i,\setminus j}\big\}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\big\{\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\big\}^2} \\ =&O_p(\sqrt{\log p/n})O_p(s\sqrt{\log p/n}) + O_p(\sqrt{s\log p/n})O_p(\sqrt{b_n}) \\ =&O_P\big(s\log p/n + \sqrt{b_n s\log p/n}\big).\end{aligned}$$

In order to have $\left|\frac{1}{n}\sum_{i=1}^{n} R_{ni,2}\right| = o_p(n^{-1/2})$ we need to have $s\log p/\sqrt{n} = o(1)$ and $b_n = o(1/\sqrt{n})$. Thus with (3) and (5) in Assumption 1, we have verified the first half condition in (C2), $\frac{1}{n}\sum_{i=1}^{n} R_{ni} = o_p(n^{-1/2})$.

Now for the second half of the condition in (C2),

$$\begin{aligned}\max_{1\le i\le n}|R_{ni,1}| &= \max_{1\le i\le n}\left|(Y_i - \mathbf{X}_i^{\intercal}\boldsymbol{\beta}^0)(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^{\intercal}\mathbf{X}_{i,\setminus j}\right| = \max_{1\le i\le n}\left|(\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^{\intercal}\mathbf{X}_{i,\setminus j}\epsilon_i\right| \\ &\le \left\|\mathbf{w}_j^0 - \hat{\mathbf{w}}_j\right\|_1 \max_{1\le i\le n}\left\|\mathbf{X}_{i,\setminus j}\epsilon_i\right\|_\infty = \left\|\mathbf{w}_j^0 - \hat{\mathbf{w}}_j\right\|_1 \max_{1\le i\le n}\max_{1\le k\le p}\left|X_{ik}\epsilon_i\right|.\end{aligned}$$

Now since $\mathbf{X}_i$ and ϵ_i are all sub-Gaussian and then we have $X_{ik}\epsilon_i$ sub-exponential, and then by the union bound, we have

$$\mathrm{P}\big(\max_{1\le i\le n}\max_{1\le k\le p}\left|X_{ik}\epsilon_i\right| > t\big) \le \sum_{1\le i\le n}\sum_{1\le k\le p}\mathrm{P}(|X_{ik}\epsilon_i| > t) \le pnC_1e^{-C_2t}.$$

By taking $t = \log(pn)/C$ with $C < C_2$, we have $\max_{1\le i\le n}\max_{1\le k\le p}\left|X_{ik}\epsilon_i\right| = O_p(\log(pn))$. Hence we have

$$\max_{1\le i\le n}|R_{ni,1}| = \left\|\mathbf{w}_j^0 - \hat{\mathbf{w}}_j\right\|_1 \max_{1\le i\le n}\max_{1\le k\le p}\left|X_{ik}\epsilon_i\right| = O_p(a_n\log(pn)).$$

In order to make $\max_{1\le i\le n}|R_{ni,1}| = o_p(n^{1/2})$, we need $a_n\log(pn)/\sqrt{n} = o(1)$, which is true under assumption (2) for a_n in Assumption 1 since $a_n\log(pn)/\sqrt{n} = o(\log(pn)/\sqrt{n\log p}) = o(\sqrt{\log p/n}) = o(1)$.

$$
\begin{aligned}
\text{Note that } \max_{1\le i\le n}|R_{ni,2}| &= \max_{1\le i\le n}|(X_{ij}-\hat{\mathbf{w}}_j^{\mathsf{T}}\mathbf{X}_{i,\setminus j})\mathbf{X}_{i,\setminus j}^{\mathsf{T}}(\boldsymbol{\beta}_{\setminus j}^0-\hat{\boldsymbol{\beta}}_{\setminus j})| \\
&\le \max_{1\le i\le n}|(X_{ij}-\mathbf{w}_j^{0\mathsf{T}}\mathbf{X}_{i,\setminus j})\mathbf{X}_{i,\setminus j}^{\mathsf{T}}(\boldsymbol{\beta}_{\setminus j}^0-\hat{\boldsymbol{\beta}}_{\setminus j})| \\
&\quad + \max_{1\le i\le n}|(\mathbf{w}_j^0-\hat{\mathbf{w}}_j)^{\mathsf{T}}\mathbf{X}_{i,\setminus j}\mathbf{X}_{i,\setminus j}^{\mathsf{T}}(\boldsymbol{\beta}_{\setminus j}^0-\hat{\boldsymbol{\beta}}_{\setminus j})| \\
&\le \|\boldsymbol{\beta}_{\setminus j}^0-\hat{\boldsymbol{\beta}}_{\setminus j})\|_1 \max_{1\le i\le n}\max_{1\le k\le p}|\eta_{ij}X_{ik}| \\
&\quad + \|(\mathbf{w}_j^0-\hat{\mathbf{w}}_j)\|_1\|(\boldsymbol{\beta}_{\setminus j}^0-\hat{\boldsymbol{\beta}}_{\setminus j})\|_1\big(\max_{1\le i\le n}\max_{1\le k\le p}|X_{ik}|\big)^2.
\end{aligned}
$$

Now since η_{ij}'s and $\mathbf{X}_i$ are all sub-Gaussian, and then by similar analysis as above we have

$$
\begin{aligned}
\max_{1\le i\le n}|R_{ni,2}| &= O_p(s\sqrt{\log p/n})O_p(\log(pn)) + O_p(a_n s\sqrt{\log p/n})O_p(\log(pn)) \\
&= O_p(s\sqrt{\log p/n}\log(pn)).
\end{aligned}
$$

In order to make $\max_{1\le i\le n}|R_{ni,2}| = o_p(n^{1/2})$, we need

$$
s\sqrt{\log p}\log(pn)/n = o_p(1),
$$

which is true under assumption (5) in Assumption 1 since $s\sqrt{\log p}\log(pn)/n = o(\sqrt{\log p/n}) = o(1)$. Thus we have $\max_{1\le i\le n}|R_{ni}| = o_p(n^{1/2})$, which verifies the second half in the condition (C2).

Now we need to check out condition (C0). From the above analysis, we have $\max_{1\le i\le n}|R_{ni}| =$

$o_p(\max_{1\le i\le n}|W_{ni}|)$. Thus we only need to prove that

$$\mathrm{P}(\min_{1\le i\le n} W_{ni} < 0 < \max_{1\le i\le n} W_{ni}) \to 1,$$

which just follows from the Gilvenko-Gantelli theorem over half-spaces as in page 219 in [Owe01]. □

Proof of Theorem 7. Notice that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R_{ni,1} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\gamma_{j1}^0 - \hat{\gamma}_{j1}) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^0(\gamma_{j1}^0 - \hat{\gamma}_{j1}) \\ + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1}).\end{aligned}$$

By condition (2) in Assumption 2 and (4.7.23) implied from condition (4) in Assumption 2,

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^0(\gamma_{j1}^0 - \hat{\gamma}_{j1})\Big| = \Big|(\gamma_{j1}^0 - \hat{\gamma}_{j1})\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \boldsymbol{\beta}^{0\intercal}\mathbf{X}_i\Big| = O_p(a_n\sqrt{\frac{\log p}{n}}),$$

and

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1})\Big| \le \Big\|\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1}\Big\|_1\Big\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathbf{X}_{i,\setminus j}\Big\|_\infty = O_p(a_n\sqrt{\frac{\log p}{n}}).$$

Thus we have

$$\frac{1}{n}\sum_{i=1}^{n} R_{ni,1} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\gamma_{j1}^0 - \hat{\gamma}_{j1}) + O_p(a_n\sqrt{\log p/n}) = O_p(a_n\sqrt{1/n}).$$

So in order to have $\frac{1}{n}\sum_{i=1}^{n} R_{ni,1} = o_p(n^{-1/2})$, we need $a_n = o_p(1)$. Note that

$$\begin{aligned}\max_{1\le i\le n}|R_{ni,1}| \le & \max_{1\le i\le n}|\epsilon_i^2(\gamma_{j1}^0 - \hat\gamma_{j1})| + \max_{1\le i\le n}|\epsilon_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^0(\gamma_{j1}^0 - \hat\gamma_{j1})| \\ & + \max_{1\le i\le n}|\epsilon_i \mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1})| \\ =&|\gamma_{j1}^0 - \hat\gamma_{j1}|\big\{\max_{1\le i\le n}|\epsilon_i^2| + \max_{1\le i\le n}|\epsilon_i \mathbf{X}_i^\intercal\boldsymbol{\beta}^0|\big\} + \|\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1}\|_1 \max_{1\le i\le n}\|\epsilon_i \mathbf{X}_{i,\setminus j}\|_\infty \\ =&|\gamma_{j1}^0 - \hat\gamma_{j1}|\big\{\max_{1\le i\le n}|\epsilon_i^2| + \max_{1\le i\le n}|\epsilon_i \mathbf{X}_i^\intercal\boldsymbol{\beta}^0|\big\} + \|\boldsymbol{\gamma}_{j,\setminus 1}^0 - \hat{\boldsymbol{\gamma}}_{j,\setminus 1}\|_1 \max_{1\le i\le n}\max_{1\le k\le p}|\epsilon_i X_{ij}|.\end{aligned}$$

And by the assumption that $\mathbf{X}_i$ and ϵ_i are sub-Gaussian, we have $\mathbf{X}_i^\intercal\boldsymbol{\beta}^0$ is sub-Gaussian and ϵ_i^2, $\epsilon_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^0$ and $X_{ij}\epsilon_i$ are all sub-exponential. Then we have

$$\mathrm{P}\big(\max_{1\le i\le n}|\epsilon_i^2| > t\big) \le n\mathrm{P}(|\epsilon_i^2| > t) \le nC_1 e^{-C_2 t}$$

which implies that $\max_{1\le i\le n}|\epsilon_i^2| = O_p(\log n)$. Thus we have $\max_{1\le i\le n}|R_{ni,1}| = O_p(a_n\log(pn))$. In order to achieve $\max_{1\le i\le n}|R_{ni,1}| = o_p(n^{1/2})$, we need $a_n\log(pn)/\sqrt{n} = o(1)$, which is true since $a_n = o(1/\sqrt{\log p})$.

$$\begin{aligned}\text{For } R_{ni,2} &= \eta_{ij,y}\mathbf{X}_i^\intercal(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) = \eta_{ij,y}\big\{X_{ij}(\beta_j^0 - \hat\beta_j) + \mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\big\} \\ &= \eta_{ij,y}\big\{[(Y_i, \mathbf{X}_{i,\setminus j}^\intercal)\boldsymbol{\gamma}_j^0 + \eta_{ij,y}](\beta_j^0 - \hat\beta_j) + \mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})\big\} \\ &= \eta_{ij,y}^2(\beta_j^0 - \hat\beta_j) + \eta_{ij,y}(Y_i, \mathbf{X}_{i,\setminus j}^\intercal)\boldsymbol{\gamma}_j^0(\beta_j^0 - \hat\beta_j) + \eta_{ij,y}\mathbf{X}_{i,\setminus j}^\intercal(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j}),\end{aligned}$$

similarly as $R_{ni,1}$, by condition (1) and (4.7.24), (4.7.25), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R_{ni,2} &= \frac{1}{n}\sum_{i=1}^{n} \eta_{ij,y}^2(\beta_j^0 - \hat{\beta}_j) + O_p(s\sqrt{\log p/n}\sqrt{\log p/n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \eta_{ij,y}^2(\beta_j^0 - \hat{\beta}_j) + O_p(s\log p/n) \\ &= O_p(s\sqrt{\log p/n}\sqrt{1/n}) + O_p(s\log p/n) = O_p(s\sqrt{\log p}/n).\end{aligned}$$

So in order to have $\frac{1}{n}\sum_{i=1}^{n} R_{ni,2} = o_p(n^{-1/2})$, we need to have $s\sqrt{\log p}/n = o_p(n^{-1/2})$, i.e. $s\sqrt{\log p/n} = o_p(1)$. Note that

$$\begin{aligned}\max_{1\le i\le n}|R_{ni,2}| \le \max_{1\le i\le n}|\eta_{ij,y}^2(\beta_j^0 - \hat{\beta}_j)| + \max_{1\le i\le n}|\eta_{ij,y}(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})\boldsymbol{\gamma}_j^0(\beta_j^0 - \hat{\beta}_j)| \\ + \max_{1\le i\le n}|\eta_{ij,y}\mathbf{X}_{i,\setminus j}^{\intercal}(\boldsymbol{\beta}_{\setminus j}^0 - \hat{\boldsymbol{\beta}}_{\setminus j})| = O_p(s\sqrt{\log p/n}\log(pn)) = o_p(\sqrt{n})\end{aligned}$$

since $s\sqrt{\log p/n}\log(pn)/\sqrt{n} = o(\sqrt{\log p/n}) = o(1)$.

Now for $R_{ni,3} = \mathbf{X}_i^{\intercal}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\{(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j)\} = (\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})^{\intercal}\mathbf{X}_i(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j)$, we have by (3) in Assumption 2

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n} R_{ni,3}\Big| &= \Big|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})^{\intercal}\mathbf{X}_i(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j)\Big| \\ &\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})^{\intercal}\mathbf{X}_i]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j)]^2} \\ &= O_p(\sqrt{s\log p/n})O_p(\sqrt{b_n}) = O_p(\sqrt{b_n s\log p/n}).\end{aligned}$$

So in order to have $\frac{1}{n}\sum_{i=1}^{n} R_{ni,3} = o_p(n^{-1/2})$, we need to have $\sqrt{b_n s\log p/n} = o_p(n^{-1/2})$,

i.e. $\sqrt{b_n s \log p} = o_p(1)$. And we also have

$$\begin{aligned}\max_{1\le i\le n} |R_{ni,3}| &\le \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_1 \|\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j\|_1 \max_{1\le i\le n}\max_{1\le j\le p} |X_{ij}| \big(\max_{1\le i\le n} |Y_i| + \max_{1\le i\le n}\max_{1\le j\le p} |X_{ij}| \big) \\ &= O_p(s\sqrt{\log p/n} a_n \log(pn)) = o_p(n^{1/2}).\end{aligned}$$

Now we need to check out condition (C0). From the above analysis, we have $\max_{1\le i\le n} |R_{ni}| = o_p(\max_{1\le i\le n} |W_{ni}|)$. Thus we only need to prove that

$$\mathrm{P}(\min_{1\le i\le n} W_{ni} < 0 < \max_{1\le i\le n} W_{ni}) \to 1,$$

which just follows from the Gilvenko-Gantelli theorem over half-spaces as in page 219 in [Owe01]. □

Proof of Theorem 8. Recall that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n R_{ni} = R_{1n} + R_{2n} + R_{3n} + R_{4n}$$

where

$$\begin{aligned}
R_{1n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n -\mathbf{X}_{i\mathcal{S}}^\intercal (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \boldsymbol{\epsilon} \big\{ X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} \big\}, \\
R_{2n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \big\{ \epsilon_i - \mathbf{X}_{i\mathcal{S}}^\intercal (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \boldsymbol{\epsilon} \big\} \big\{ \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbf{X}_{i\mathcal{S}} \big\}, \\
R_{3n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \big\{ X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} \big\} \big\{ \mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}^*} \big\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}], \\
R_{4n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \big\{ \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbf{X}_{i\mathcal{S}} \big\} \\
&\qquad \times \big\{ \mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}^*} \big\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}].
\end{aligned}$$

Now for R_{1n}, we have

$$\begin{aligned} R_{1n} &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} \mathbf{X}_{i\mathcal{S}}^{\intercal} (\mathbb{X}_{\mathcal{S}}^{\intercal} \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^{\intercal} \boldsymbol{\epsilon} \\ &= -\Big\{ \frac{1}{n} \sum_{i=1}^{n} \{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} \mathbf{X}_{i\mathcal{S}}^{\intercal} \Big\} \Big\{ \sqrt{n} (\mathbb{X}_{\mathcal{S}}^{\intercal} \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^{\intercal} \boldsymbol{\epsilon} \Big\}. \end{aligned}$$

Now we need to bound the two terms $\frac{1}{n} \sum_{i=1}^{n} \{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} \mathbf{X}_{i\mathcal{S}}$ and $\sqrt{n} (\mathbb{X}_{\mathcal{S}}^{\intercal} \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^{\intercal} \boldsymbol{\epsilon}$. In fact, for every $k \in \mathcal{S}$, we have that the two Gaussian random variables $X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}$ and X_{ik} have the following properties:

$$\begin{aligned} &\mathrm{E}(X_{ik}) = \mathrm{E}(X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}) = 0; \\ &\mathrm{E}(X_{ik}^2) = \sigma_{kk}, \ \mathrm{E}[(X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}})^2] = \sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{\Sigma}_{\mathcal{S}j}; \\ &\mathrm{Cov}(X_{ik}, X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}) = \mathrm{E}[X_{ik}(X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}})] = \sigma_{kj} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{\Sigma}_{\mathcal{S}k} \\ &\qquad\qquad = \sigma_{kj} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}} \mathbf{e}_k = \sigma_{kj} - \boldsymbol{\Sigma}_{j\mathcal{S}} \mathbf{e}_k = \sigma_{kj} - \sigma_{jk} = 0. \end{aligned}$$

Thus we have

$$\begin{pmatrix} X_{ik} \\ X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} \end{pmatrix} \sim \mathrm{N}\Big(\mathbf{0}, \begin{pmatrix} \sigma_{kk} & 0 \\ 0 & \sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{\Sigma}_{\mathcal{S}j} \end{pmatrix} \Big). \tag{4.7.35}$$

Under (1) in Assumption 3, by Lemma A.3 from [BL08], we have there exists constants $C, C_1, C_2 > 0$ such that

$$\mathrm{P}\big\{ \big| \frac{1}{n} \sum_{i=1}^{n} \{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} X_{ij} \big| \geq t \big\} \leq C_1 \exp(-C_2 n t^2), \text{ for } 0 \leq t \leq C.$$

By union inequality, we then have

$$\mathrm{P}\big\{\max_{\mathcal{S}:|\mathcal{S}|\leq m}\big\|\frac{1}{n}\sum_{i=1}^{n}\big\{X_{ij}-\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\big\}\mathbf{X}_{i\mathcal{S}}\big\|_{\infty}\geq t\big\}\leq C_1 mp^m\exp(-C_2nt^2),$$

for $0\leq t\leq C$, where $|\{\mathcal{S}\subseteq\{1,2,,\cdots,p\}:|\mathcal{S}|\leq m\}|\leq p^m$.

For $mp^m\exp(-C_2nt^2)=\exp(-C_2nt^2+m\log p+\log m)$, take

$$t=\sqrt{\frac{m\log p+\log m+C\log p}{(C_2n)}}\sim\sqrt{m\log p/n},$$

and then we have

$$\max_{\mathcal{S}:|\mathcal{S}|\leq m}\big\|\frac{1}{n}\sum_{i=1}^{n}\big\{X_{ij}-\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\big\}\mathbf{X}_{i\mathcal{S}}\big\|_{\infty}=O_p(\sqrt{m\log p/n}).$$

Now in order to control $\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}$, first notice that by the following matrix equality [HS81]

$$\begin{aligned}&(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n)^{-1}=\big\{\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}+(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n-\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})\big\}^{-1}\\ =&\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}-\underbrace{\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\big\{\mathbf{I}+(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n-\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\big\}^{-1}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n-\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}}_{\boldsymbol{\Delta}_{\mathcal{S}}},\end{aligned}\tag{4.7.36}$$

we have

$$\begin{aligned}\|\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\|_1 &= \|(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n)^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_1 \\ &\leq \|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_1 + \|\boldsymbol{\Delta}_{\mathcal{S}}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_1 \\ &\leq \sqrt{|\mathcal{S}|}\|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 + \sqrt{|\mathcal{S}|}\|\boldsymbol{\Delta}_{\mathcal{S}}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 \\ &\leq \sqrt{|\mathcal{S}|}\|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 + \sqrt{|\mathcal{S}|}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2.\end{aligned}$$

One of the most important results in matrix analysis is the Cauchy (eigenvalue) interlacing theorem. It asserts that the eigenvalues of any principal submatrix of a symmetric matrix interlace those of the symmetric matrix. For example, if an $n \times n$ symmetric matrix $\mathbf{S}$ can be partitioned as

$$\mathbf{S} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{C} \end{pmatrix},$$

in which $\mathbf{A}$ is an $r \times r$ principle submatrix, then for each $i \in 1, 2, \cdots, r$, we have

$$\lambda_i(\mathbf{S}) \leq \lambda_i(\mathbf{A}) \leq \lambda_{n-r+i}(\mathbf{S}).$$

In particular, we have $\lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})$ and $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})$. Thus by the definition of maximum eigenvalue, we have

$$\|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 \leq \lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2.$$

So

$$\|\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\|_1 \leq \sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 + \sqrt{|\mathcal{S}|}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2$$
$$= \sqrt{|\mathcal{S}|}\{\lambda_{\min}^{-1} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_2\}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2.$$

Now we have to control $\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2$ and $\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2$. In order to control the first one, by the sub-Gaussian tailed condition (2) in Assumption 3,

$$\mathrm{P}(\max_{\mathcal{S}:|\mathcal{S}|\leq m}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 \geq t\sqrt{n}) \leq \mathrm{P}(\max_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{j\in\mathcal{S}}|\frac{1}{n}\sum_{i=1}^{n}X_{ij}\epsilon_i| \geq t/\sqrt{m})$$
$$\leq p^m m\exp(-Cnt^2/m),$$

followed from the Bernstein inequality for t small. For $p^m m\exp(-Cnt^2/m) = \exp(m\log p + \log m - Cnt^2/m)$, take $t = \sqrt{m}\sqrt{\frac{m\log p+\log m+C_1\log p}{Cn}} \sim \sqrt{m^2\log p/n}$. Then we have the following order

$$\max_{\mathcal{S}:|\mathcal{S}|\leq m}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 = O_p(m\sqrt{\log p}).$$

Now for $\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2$ with $\boldsymbol{\Delta}_{\mathcal{S}} = \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\{\mathbf{I} + (\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\}^{-1}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}$, we have to control $\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}$ first. Note that

$$\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}\|_2 \geq \epsilon\Big) \leq \mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{j,k}|\mathbb{X}_j^{\intercal}\mathbb{X}_k/n - \sigma_{jk}| \geq \epsilon/m\Big)$$
$$\leq m^2p^m\mathrm{P}\Big(|\mathbb{X}_j^{\intercal}\mathbb{X}_k/n - \sigma_{jk}| \geq \epsilon/m\Big) \leq C_1m^2p^m\exp(-C_2n\epsilon^2/m^2)$$

where the last inequality is also followed from Lemma A.3 in [BL08] with constants $C_1, C_2 > 0$. For $m^2p^m\exp(-C_2n\epsilon^2/m^2) = \exp(2\log m + m\log p - C_2n\epsilon^2/m^2)$, by taking $\epsilon = m\sqrt{\frac{m\log p+2\log m+C_1\log}{C_2n}}$

$\sqrt{m^3 \log p/n}$, we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}\|_2 = O_p(\sqrt{m^3 \log p/n}).$$

It follows then

$$\begin{aligned}
\|\mathbf{\Delta}_{\mathcal{S}}\|_2 &= \|\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\{\mathbf{I} + (\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\}^{-1}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_2 \\
&\leq \|\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_2^2 \|\mathbf{I} + (\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\}^{-1}\|_2 \|\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}\|_2 \\
&= O_p(\sqrt{m^3 \log p/n}),
\end{aligned}$$

since $\|\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_2 = \lambda_{\max}^{1/2}(\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-2}) \leq \lambda_{\min}^{-1}$.

Thus we have

$$\begin{aligned}
\|\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\|_1 &\leq \sqrt{|\mathcal{S}|}\{\lambda_{\min}^{-1} + \|\mathbf{\Delta}_{\mathcal{S}}\|_2\}\|\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\|_2 \\
&= O_p(\sqrt{m^3 \log p/n}),
\end{aligned}$$

i.e. $\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\|_1 = O_p(\sqrt{m^3 \log p/n})$.

In summary, we then have

$$\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \left|\left\{\frac{1}{n}\sum_{i=1}^{n}\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}\mathbf{X}_{i\mathcal{S}}^{\intercal}\right\}\left\{\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\right\}\right| \\
\leq &\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \left\|\frac{1}{n}\sum_{i=1}^{n}\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}\mathbf{X}_{i\mathcal{S}}^{\intercal}\right\|_\infty \sup_{\mathcal{S}:|\mathcal{S}|\leq m} \left\|\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\right\|_1 \\
= &O_p(\sqrt{m \log p/n})O_p(\sqrt{m^3 \log p/n}) = O_p(m^2 \log p/n).
\end{aligned}$$

And hence $R_{1n} = o_p(1)$.

For R_{2n}, we have

$$
\begin{aligned}
R_{2n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\epsilon_i - \mathbf{X}_{i\mathcal{S}}^{\intercal}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\}\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^{\intercal}\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbf{X}_{i\mathcal{S}}\} \\
&= \{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbb{X}_j^{\intercal}\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\mathbf{X}_{i\mathcal{S}}\epsilon_i - \mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^{\intercal}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\} \\
&= \{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbb{X}_j^{\intercal}\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\}\Big\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_{i\mathcal{S}}\epsilon_i - \{\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^{\intercal}\}\sqrt{n}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}\Big\} \\
&= \{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbb{X}_j^{\intercal}\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\}\Big\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_{i\mathcal{S}}\epsilon_i - \mathbb{X}_{\mathcal{S}}^{\intercal}\boldsymbol{\epsilon}/\sqrt{n}\Big\} = 0.
\end{aligned}
$$

Observe that we can rewrite R_{3n} as

$$
\begin{aligned}
R_{3n} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}\{\mathbf{X}_{i\mathcal{S}^*}^{\intercal} - \mathbf{X}_{i\mathcal{S}}^{\intercal}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}^*}\}[\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] \\
=& \frac{1}{\sqrt{n}}\mathbb{X}_j^{\intercal}\{\mathbf{I} - \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^{\intercal}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^{\intercal}\}\mathbb{X}_{\mathcal{S}^*}[\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}],
\end{aligned}
$$

where $\frac{1}{\sqrt{n}}\mathbb{X}_j^\intercal\{\mathbf{I}-\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\}\mathbb{X}_{\mathcal{S}^*}$ can be controlled as follows

$$
\begin{aligned}
&\|\frac{1}{\sqrt{n}}\mathbb{X}_j^\intercal\{\mathbf{I}-\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\}\mathbb{X}_{\mathcal{S}^*}\|_\infty = \max_{k\in\mathcal{S}^*}|\frac{1}{\sqrt{n}}\mathbb{X}_j^\intercal\{\mathbf{I}-\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\}\mathbb{X}_k| \\
\leq&\sqrt{n}\max_{k\in\mathcal{S}^*}\Big\{\big|\mathbb{X}_j^\intercal\mathbb{X}_k/n-\sigma_{jk}\big|+\big|\sigma_{jk}-\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}\big|+\big|[\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}]\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}\big| \\
&+\big|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}[\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}]\big|+\big|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}k}\big| \\
&+\big|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}}[\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}]\big|+\big|[\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}]\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}[\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}]\big| \\
&+\big|[\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}]\mathbf{\Delta}_{\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}k}\big|+\big|[\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}]\mathbf{\Delta}_{\mathcal{S}}[\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}]\big|\Big\} \\
\leq&\sqrt{n}\max_{k\in\mathcal{S}^*}\Big\{\big|\mathbb{X}_j^\intercal\mathbb{X}_k/n-\sigma_{jk}\big|+\big|\sigma_{jk}-\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}\big|+\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}\|_\infty\sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max} \\
&+\sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max}\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}\|_\infty+\lambda_{\max}^2\|\mathbf{\Delta}_{\mathcal{S}}\|_2 \\
&+\sqrt{|\mathcal{S}|}\lambda_{\max}\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}\|_\infty+\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}\|_2\lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}\|_2 \\
&+\sqrt{|\mathcal{S}|}\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}\|_\infty\|\mathbf{\Delta}_{\mathcal{S}}\|_2\lambda_{\max}+\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}\|_2\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n-\mathbf{\Sigma}_{\mathcal{S}k}\|_2\Big\}.
\end{aligned}
$$

And we have that

$$
\begin{aligned}
&\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{k\in\mathcal{S}^*}|\sigma_{jk}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_k|\geq\epsilon\Big) \\
\leq&p^{m+1}\mathrm{P}\Big(|\sigma_{jk}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_k|\geq\epsilon\Big)\leq C_1p^{m+1}\exp(-C_2n\epsilon^2)
\end{aligned}
$$

where the last inequality is also followed from Lemma A.3 in [BL08] with constants $C_1, C_2 > 0$. For $p^{m+1}\exp(-C_2n\epsilon^2)=\exp((m+1)\log p-C_2n\epsilon^2)$, by taking $\epsilon=\sqrt{\frac{(m+1)\log p+C_1\log p}{C_2n}}\sim\sqrt{m\log p/n}$, we have

$$
\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{k\in\mathcal{S}^*}|\sigma_{jk}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_k|=O_p(\sqrt{m\log p/n}).
$$

Similarly, we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\mathbf{\Sigma}_{j\mathcal{S}} - \frac{1}{n}\mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}\|_\infty = O_p(\sqrt{m\log p/n})$$

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \max_{k\in\mathcal{S}^*} \|\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_k/n - \mathbf{\Sigma}_{\mathcal{S}k}\|_\infty = O_p(\sqrt{m\log p/n})$$

By $\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\mathbf{\Delta}_{\mathcal{S}}\|_2 = O_p(\sqrt{m^3\log p/n})$, we have

$$\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\frac{1}{\sqrt{n}}\mathbb{X}_j^\intercal\{\mathbf{I} - \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\}\mathbb{X}_{\mathcal{S}^*}\|_\infty \\
\leq&\sqrt{n}\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{k\in\mathcal{S}^*}\Big\{|\mathbb{X}_j^\intercal\mathbb{X}_k/n - \sigma_{jk}| + |\sigma_{jk} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}| \\
&+ \|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{j\mathcal{S}}\|_\infty\sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max} \\
&+ \sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max}\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \mathbf{\Sigma}_{\mathcal{S}k}\|_\infty + \lambda_{\max}^2\|\mathbf{\Delta}_{\mathcal{S}}\|_2 \\
&+ \sqrt{|\mathcal{S}|}\lambda_{\max}\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \mathbf{\Sigma}_{\mathcal{S}k}\|_\infty + \sqrt{|\mathcal{S}|}\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{j\mathcal{S}}\|_\infty\|\mathbf{\Delta}_{\mathcal{S}}\|_2\lambda_{\max} \\
&+ \|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{j\mathcal{S}}\|_2\lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \mathbf{\Sigma}_{\mathcal{S}k}\|_2 \\
&+ \|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n - \mathbf{\Sigma}_{j\mathcal{S}}\|_2\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \mathbf{\Sigma}_{\mathcal{S}k}\|_2\Big\} \\
=&\sqrt{n}\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{k\in\mathcal{S}^*}|\sigma_{jk} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}| + O_p\{\sqrt{n}\sqrt{m^3\log p/n}\},
\end{aligned}$$

since $\sqrt{m^3\log p/n} = o(1)$. Under condition (4) and (5) in Assumption 3, we have that

$R_{3n} = o_p(1)$.

$$\begin{aligned}
\text{Note that } R_{4n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbf{X}_{i\mathcal{S}}\} \\
&\qquad \times \{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}^*}\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}} \mathbf{X}_{i\mathcal{S}^*}^\intercal\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] \\
&\quad - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}^*}/\sqrt{n}) [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbf{X}_{i\mathcal{S}} \mathbf{X}_{i\mathcal{S}^*}^\intercal\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] \\
&\quad + \{\mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}^*}/\sqrt{n}\} [\boldsymbol{\beta}_{\mathcal{S}^*}^0 - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}] = 0.
\end{aligned}$$

Thus we have verified that $\frac{1}{n}\sum_{i=1}^{n} R_{ni} = o_p(n^{-1/2})$.

And for $R_{ni,1}$, we have

$$\begin{aligned}
\max_{1\le i\le n} |R_{ni,1}| &= \|(\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \boldsymbol{\epsilon}\|_1 \max_{1\le i\le n} \|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} \mathbf{X}_{i\mathcal{S}}^\intercal\|_\infty \\
&= \|(\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \boldsymbol{\epsilon}\|_1 \max_{1\le i\le n} \max_{k\in\mathcal{S}} \left|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} X_{ik}\right|
\end{aligned}$$

where $\sup_{\mathcal{S}:|\mathcal{S}|\le m} \|(\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}^\intercal \boldsymbol{\epsilon}\|_1 = O_p(\sqrt{m^3 \log p}/n)$. And since $X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}$ is Gaussian under the assumption that $\mathbf{X}$ is Gaussian, we have $\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} X_{ik}$ sub-exponential. So

$$\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{k\in\mathcal{S}} \left|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} X_{ik}\right| > t\Big) \le p^m nm C_1 \exp(-C_2 t)$$

which leads to $\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{k\in\mathcal{S}} \left|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{X}_{i\mathcal{S}}\} X_{ik}\right| = O_p(m \log p)$.

Thus we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \max_{1\leq i\leq n} |R_{ni,1}| = O_p(m\log p\sqrt{m^3\log p}/n) = o_p(n^{1/2})$$

since $(m\log p/n)\sqrt{m^3\log p/n} = o(1)$.

And for $R_{ni,2}$, we have

$$\max_{1\leq i\leq n} |R_{ni,2}| \leq \|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\|_1 \max_{1\leq i\leq n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i - \mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_\infty,$$

where

$$\begin{aligned}
&\|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\|_1 = \|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}}/n)^{-1}\|_1\\
=&\|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1} - \mathbf{\Delta}_{\mathcal{S}})\|_1\\
\leq&\|(\mathbf{\Sigma}_{j\mathcal{S}} - n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 + \|n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}}\|_1\\
\leq&\|(\mathbf{\Sigma}_{j\mathcal{S}} - n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 + \|(n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}} - \mathbf{\Sigma}_{j\mathcal{S}})\mathbf{\Delta}_{\mathcal{S}}\|_1 + \|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}}\|_1.
\end{aligned}$$

And by simple algebra, we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|(\mathbf{\Sigma}_{j\mathcal{S}} - n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 = O_p(\sqrt{m^3\log p/n}),$$

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|(n^{-1}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}} - \mathbf{\Sigma}_{j\mathcal{S}})\mathbf{\Delta}_{\mathcal{S}}\|_1 = O_p(m^2\log p/n),$$

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Delta}_{\mathcal{S}}\|_1 = O_p(m^2\sqrt{\log p/n}).$$

Now for

$$\begin{aligned}
&\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i - \mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_\infty \\
\le &\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i\|_\infty + \max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_\infty \\
\le &\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i\|_\infty + \|(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_\infty \max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal\|_\infty,
\end{aligned}$$

since $X_{ik}\epsilon_i$ is sub-exponential, we have

$$\begin{aligned}
\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i\|_\infty > t\Big) &= \mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n}\max_{k\in\mathcal{S}} |X_{ik}\epsilon_i| > t\Big) \\
&\le p^m mnC_1e^{-C_2t}
\end{aligned}$$

which leads to $\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i\|_\infty = O_p(m\log p)$. And since $X_{ik}X_{il}$ is sub-exponential, we have

$$\begin{aligned}
\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal\|_\infty > t\Big) &\le \mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n}\sqrt{m}\max_{k,l\in\mathcal{S}} |X_{ik}X_{il}| > t\Big) \\
&\le p^m m^2 nC_1e^{-C_2t}
\end{aligned}$$

which leads to $\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal\|_\infty = O_p(\sqrt{m}m\log p)$.

Since $\sup_{\mathcal{S}:|\mathcal{S}|\le m} \|(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_1 = O_p(\sqrt{m^3\log p}/n)$, we have

$$\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\mathbf{X}_{i\mathcal{S}}\epsilon_i - \mathbf{X}_{i\mathcal{S}}\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\|_\infty \\
=&O_p(m\log p + \sqrt{m}m\log p\sqrt{m^3\log p}/n) = O_p(m\log p(1 + m^2\sqrt{\log p}/n)).
\end{aligned}$$

In summary,

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n}|R_{ni,2}| = O_p\{m^3\log p\sqrt{\log p/n}(1+m^2\sqrt{\log p}/n)\},$$

since $\log p/n \to 0$. In order to have $\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n}|R_{ni,2}| = o_p(n^{1/2})$, we need to have $m^3(\log p/\sqrt{n})\sqrt{\log p/n} = o(1)$, which is true under (4) in Assumption 3 since $m^3(\log p/\sqrt{n})\sqrt{\log p/n} = \sqrt{m^3\log p/n}\sqrt{(\log p)^2m^3/n} = o(1)$.

$$\begin{aligned}\text{Observe that } \max_{1\le i\le n}|R_{ni,3}| &\le \|\boldsymbol{\beta}^0_{\mathcal{S}^*} - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}\|_1 \\ &\times \max_{1\le i\le n}\|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}^*}\}\|_\infty.\end{aligned}$$

Since

$$\begin{aligned}\|\mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}^*}\|_\infty \le \max_{k\in\mathcal{S}^*}\big\{&|\mathbf{X}^\intercal_{i\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}k}| + |\mathbf{X}^\intercal_{i\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})| \\ &+ |\mathbf{X}^\intercal_{i\mathcal{S}}\boldsymbol{\Delta}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}k}| + |\mathbf{X}^\intercal_{i\mathcal{S}}\boldsymbol{\Delta}_{\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})|\big\},\end{aligned} \tag{4.7.37}$$

we have

$$\begin{aligned}&\max_{1\le i\le n}\|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\ \le &\max_{1\le i\le n}\max_{k\in\mathcal{S}^*}|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}X_{ik}| + \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}\mathbf{X}^\intercal_{i\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}k}| \\ &+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}}|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}X_{i\mathcal{S}}\}X_{il}|\|\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})\|_1 \\ &+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}}|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}X_{il}|\sqrt{m}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2\|\boldsymbol{\Sigma}_{\mathcal{S}k}\|_2 \\ &+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}}|\{X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\}X_{il}|\|\boldsymbol{\Delta}_{\mathcal{S}}(\mathbb{X}^\intercal_{\mathcal{S}}\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})\|_1.\end{aligned}$$

Now since

$$\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{k\in\mathcal{S}^*} |\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}X_{ik}| > t\Big) \le p^{m+1} n C_1 e^{-C_2 t},$$

we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{k\in\mathcal{S}^*} |\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}X_{ik}| = O_p(m\log p).$$

Similarly, we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{k\in\mathcal{S}^*} |\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}\mathbf{X}_{i\mathcal{S}}^\intercal\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}| = O_p(m\log p),$$

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \max_{l\in\mathcal{S}} |\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}X_{il}| = O_p(m\log p).$$

And then by simple algebra, we have

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{k\in\mathcal{S}^*} \|(\mathbf{\Sigma}_{k\mathcal{S}} - n^{-1}\mathbb{X}_k^\intercal\mathbb{X}_{\mathcal{S}})\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 = O_p(\sqrt{m^3\log p/n}),$$

$$\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{k\in\mathcal{S}^*} \|(n^{-1}\mathbb{X}_k^\intercal\mathbb{X}_{\mathcal{S}} - \mathbf{\Sigma}_{k\mathcal{S}})\mathbf{\Delta}_{\mathcal{S}}\|_1 = O_p(m^2\log p/n).$$

Thus we have

$$\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\le m} \max_{1\le i\le n} \|\{X_{ij} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\}\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
=&O_p\{m\log p(1+\sqrt{m^3\log p/n} + m^2\sqrt{\log p/n} + m^2\log p/n)\} \\
=&O_p\{m\log p(1+\sqrt{m^3\log p/n} + m^2\log p/n)\},
\end{aligned}$$

which leads to

$$\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \max_{1\leq i\leq n} |R_{ni,3}| = O_p(s\sqrt{\log p/n}m\log p(1+\sqrt{m^3\log p/n}+m^2\log p/n)).$$

In order to have $\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\max_{1\leq i\leq n}|R_{ni,3}| = o_p(n^{1/2})$, we need

$$s\sqrt{\log p/n}(m\log p/\sqrt{n})(1+\sqrt{m^3\log p/n}+m^2\log p/n) = o(1),$$

which is true under (4) in Assumption 3.

And for $\max_{1\leq i\leq n}|R_{ni,4}| = \|\boldsymbol{\beta}^0_{\mathcal{S}^*} - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}\|_1$

$$\times \max_{1\leq i\leq n} \|\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty.$$

And for

$$\begin{aligned}
&\max_{1\leq i\leq n} \|\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n(\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}} - \boldsymbol{\Delta}_{\mathcal{S}})\}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
=&\max_{1\leq i\leq n} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
&+\max_{1\leq i\leq n} \|(\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{j\mathcal{S}})\boldsymbol{\Delta}_{\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
&+\max_{1\leq i\leq n} \|\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Delta}_{\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}^\intercal_{i\mathcal{S}^*} - \mathbf{X}^\intercal_{i\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty,
\end{aligned}$$

by (4.7.37), we have

$$
\begin{aligned}
&\max_{1\le i\le n} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
\le &\max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 |X_{il}X_{ik}| \\
&+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 X_{il}^2 \|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}k}\|_1 \\
&+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 X_{il}^2 \|\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})\|_1 \\
&+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 X_{il}^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}k}\|_1 \\
&+ \max_{1\le i\le n}\max_{k\in\mathcal{S}^*}\max_{l\in\mathcal{S}} \|(\boldsymbol{\Sigma}_{j\mathcal{S}} - \mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n)\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_1 X_{il}^2 \|\boldsymbol{\Delta}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_k/n - \boldsymbol{\Sigma}_{\mathcal{S}k})\|_1 \\
&= O_p(m^3\log p\sqrt{\log p/n}),
\end{aligned}
$$

under the condition that $m^3\log p/n \to 0$. Similarly we have

$$
\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|(\mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{j\mathcal{S}})\boldsymbol{\Delta}_{\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
&\qquad = O_p\{m^{7/2}(\log p)^2/n\} \\
&\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} \|\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Delta}_{\mathcal{S}}\mathbf{X}_{i\mathcal{S}}\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\}\|_\infty \\
&\qquad = O_p\{m^{7/2}\log p\sqrt{\log p/n}\}
\end{aligned}
$$

if $m^3\log p/n \to 0$. In summary, if $m^3\log p/n \to 0$,

$$
\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{1\le i\le n} |R_{ni,4}| = O_p\{sm^{7/2}(\log p)^2/n\}.
$$

Thus in order to have $\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \max_{1\leq i\leq n} |R_{ni,4}| = o_p(n^{1/2})$, we need

$$sm^{7/2}(\log p)^2/n^{3/2} = o(1),$$

which is true under the condition (4) in Assumption 3 since $sm^{7/2}(\log p)^2/n^{3/2} = s\sqrt{\frac{(\log p)^4 m^7}{n^3}} = s\sqrt{\frac{(\log p)^2 m^3}{n}} m^2 \log p/n = o(1)$.

From the above analysis, we have $\max_{1\leq i\leq n} |R_{ni}| = o_p(\max_{1\leq i\leq n} |W_{ni}|)$. Thus we only need to prove that

$$\mathrm{P}(\min_{1\leq i\leq n} W_{ni} < 0 < \max_{1\leq i\leq n} W_{ni}) \to 1,$$

which just follows from the Gilvenko-Gantelli theorem over half-spaces as in page 219 in [Owe01]. □

For the proof of the three propositions, they are just followed from the proof of the corresponding theorems. We here just prove the Proposition 2.

Proof of Proposition 6. In order to get the asymptotic normality of $\hat{\beta}_j^{(\text{kfc-de})}$, we have to deal with $\frac{1}{n}\sum_{i=1}^n \tilde{X}_{ij}^2$. Now since

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \tilde{X}_{ij}^2 &= \frac{1}{n}\sum_{i=1}^n \left\{X_{ij} - \mathbb{X}_j^\intercal \mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S})^{-1}\mathbf{X}_{i\mathcal{S}}\right\}^2 \\ &= \frac{1}{n}\mathbb{X}_j^\intercal \mathbb{X}_j - \frac{1}{n}\mathbb{X}_j^\intercal \mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S})^{-1}\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_j = \frac{1}{n}\mathbb{X}_j^\intercal \{\mathbf{I} - \mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S})^{-1}\mathbb{X}_\mathcal{S}^\intercal\}\mathbb{X}_j,\end{aligned}$$

we have

$$
\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}^2-(\sigma_{jj}-\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})|\\
=&|\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_j-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}}/n)^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_j/n-(\sigma_{jj}-\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})|\\
\leq&\Big\{\big|\mathbb{X}_j^\intercal\mathbb{X}_j/n-\sigma_{jj}\big|+2\|\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}/n-\mathbf{\Sigma}_{j\mathcal{S}}\|_\infty\sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max}\\
&+\lambda_{\max}^2\|\mathbf{\Delta}_{\mathcal{S}}\|_2+2\sqrt{|\mathcal{S}|}\lambda_{\max}\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_j/n-\mathbf{\Sigma}_{\mathcal{S}j}\|_\infty\\
&+\lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_j/n-\mathbf{\Sigma}_{\mathcal{S}j}\|_2^2+\|\mathbf{\Delta}_{\mathcal{S}}\|_2\|\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_j/n-\mathbf{\Sigma}_{\mathcal{S}j}\|_2^2\Big\}.
\end{aligned}
$$

And since

$$
\begin{aligned}
&\mathrm{P}\Big(\sup_{\mathcal{S}:|\mathcal{S}|\leq m}|\sigma_{jj}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_j|\geq\epsilon\Big)\\
\leq&p^m\mathrm{P}\Big(|\sigma_{jj}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_j|\geq\epsilon\Big)\leq C_1p^m\exp(-C_2n\epsilon^2)
\end{aligned}
$$

we have

$$
\sup_{\mathcal{S}:|\mathcal{S}|\leq m}|\sigma_{jj}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_j|=O_p(\sqrt{m\log p/n}).
$$

Now for the term $\|\mathbf{\Sigma}_{j\mathcal{S}}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}\|_\infty$, we have proved above that

$$
\sup_{\mathcal{S}:|\mathcal{S}|\leq m}\|\big(\mathbf{\Sigma}_{j\mathcal{S}}-\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}\big)\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\|_\infty=O_p(\sqrt{m\log p/n}).
$$

By $\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \|\boldsymbol{\Delta}_{\mathcal{S}}\|_2 = O_p(\sqrt{m^3 \log p/n})$, we have

$$
\begin{aligned}
&\sup_{\mathcal{S}:|\mathcal{S}|\leq m} |\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_{ij}^2 - (\sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j})| \\
\leq &\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \Big\{ \big|\mathbb{X}_j^\intercal \mathbb{X}_j/n - \sigma_{jj}\big| + 2\|\mathbb{X}_j^\intercal \mathbb{X}_{\mathcal{S}}/n - \boldsymbol{\Sigma}_{j\mathcal{S}}\|_\infty \sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max} \\
&+ \lambda_{\max}^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_2 + 2\sqrt{|\mathcal{S}|}\lambda_{\max}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2 \|\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_j/n - \boldsymbol{\Sigma}_{\mathcal{S}j}\|_\infty \\
&+ \lambda_{\min}^{-1}\|\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_j/n - \boldsymbol{\Sigma}_{\mathcal{S}j}\|_2^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_2 \|\mathbb{X}_{\mathcal{S}}^\intercal \mathbb{X}_j/n - \boldsymbol{\Sigma}_{\mathcal{S}j}\|_2^2 \Big\} \\
= &\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \Big\{ O_p(\sqrt{m\log p/n}) + O_p(\sqrt{m\log p/n})\sqrt{|\mathcal{S}|}\lambda_{\min}^{-1}\lambda_{\max} \\
&+ \lambda_{\max}^2 O_p(\sqrt{m^3\log p/n}) + \sqrt{|\mathcal{S}|}\lambda_{\max} O_p(\sqrt{m^3\log p/n}) O_p(\sqrt{m\log p/n}) \\
&+ |\mathcal{S}| O_p(\sqrt{m\log p/n})^2 \lambda_{\min}^{-1} + |\mathcal{S}| O_p(\sqrt{m\log p/n})^2 O_p(\sqrt{m^3\log p/n}) \Big\} \\
= &O_p\{\sqrt{m^3\log p/n}\}.
\end{aligned}
$$

Thus we have

$$
\sup_{\mathcal{S}:|\mathcal{S}|\leq m} \Big|\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_{ij}^2 - (\sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j})\Big| = O_p\{\sqrt{m^3\log p/n}\} = o_p(1). \tag{4.7.38}
$$

Hence we have the following asymptotic normality by Slutsky's theorem

$$
\sqrt{n}(\hat{\beta}_j^{(\text{kfc-de})} - \beta_j^0) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_{ni}^{(\text{kfc})}(\beta_j^0)}{\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}^2} \xrightarrow{d} \mathrm{N}(0, \sigma_{\text{kfc}}^2),
$$

where $\sigma_{\text{kfc}}^2 = \lim_{n\to\infty}(\theta_{jj} - 2\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Theta}_{j\mathcal{S}} + \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j})/(\sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j})$.

□

Chapter 5

Conclusions and Future Directions

In this chapter, we aim to reiterate the main contributions of this thesis, and to outline some of the things that could possibly follow as future developments on the results presented here. In Section 5.1, we start with the summary of the main ideas in the thesis, especially from Chapters 2, 3 and 4. Section 5.2 layouts some natural extensions of the ideas in this thesis.

5.1 Summary and Contributions

In Chapter 2 and 3, we proposed EL based procedures to make pointwise and simultaneous inferences on functional linear models, treating sparse and dense functional data in a unified framework. We showed that EL is a nice tool to accomplish this goal. We studied the asymptotic distributions of the EL based test statistics under the null and local alternative hypotheses for both sparse and dense functional data. We established the transition phase in η, the order of repeated measurements, for pointwise and simultaneous tests. The transition point η_0 was shown to be 1/8 for the pointwise test and 1/16 for the simultaneous test. If $\eta \leq \eta_0$, we showed that the proposed method is able to detect alternatives of size $b_n^* = n^{-4(1+\eta)/9}$ for the pointwise test and of order $b_n^* = n^{-8(1+\eta)/17}$ for the simultaneous test. For dense functional data such that $\eta > \eta_0$, we found that the proposed tests are able to detect alternatives of magnitude $n^{-1/2}$ both pointwisely and simultaneously, which is the same order of alternative a parametric test can detect. Moreover, we proposed a practical

bandwidth selection method for functional data. Many bandwidth selection methods were proposed for independent or weakly dependent data, but bandwidth selection for functional data remained a challenging problem, see [ZPW13] for a recent study. Numerical experiments in Chapter 2 showed that the proposed bandwidth selection method works well in practice.

In Chapter 4, we proposed a unified framework for high dimensional inference based on the empirical likelihood which is constructed with estimating equations. It can be used to test statistical hypothesis and construct confidence intervals, which have more natural data driven shape. To broaden the applicability of the method, the general theory was presented with the general conditions to be satisfied. In principal, all of the methods proposed in the existing literature can be re-considered under our framework and make fair comparison among them, although the technical details can be different case by case. Moreover, the key advantage of our proposed likelihood ratio based method comparing with others such as Wald type method and Score based method is that it can allow heteroscedastic error noise. This is largely due to the nice self normalization property of the empirical likelihood formulation. In particular, we did not assume independence between the error term and the covariates, which is a common assumption in the existing literature, although we made the uncorrelatedness assumption.

5.2 Future Directions

This thesis focused on applying empirical likelihood to solve some fundamental problems in simple statistical models, especially linear models. Hence a natural direction for future research is to generalize our methodologies to more complicated statistical models, such as generalized linear models and survival models. For functional linear models in Chapter 2

and 3, we gained the robustness in terms of the correlation structure of the error process. But if we have prior knowledge of the error process, how to incorporate the error correlation information into the estimation and inference procedures to increase the efficiency is a very interesting topic for future investigation. We only considered one general type of hypothesis in Chapter 2 and 3. There is another hypothesis problem, goodness of fit testing, which could be another promising research problem. For the high dimensional linear model in Chapter 4, we only focused on one estimating equation. But when we have more than one estimating equations, how to combine all of the estimating equations to make more efficient inference is worthy of further investigation. In general, the self-normalization property of EL is powerful and we should make use of it to solve some problems in various statistical analysis.

BIBLIOGRAPHY

BIBLIOGRAPHY

[AS58] J. Aitchison and S.D. Silvey, *Maximum-likelihood estimation of parameters subject to restraints*, The Annals of Mathematical Statistics **29** (1958), no. 3, 813–828.

[B+13] Peter Bühlmann et al., *Statistical significance in high-dimensional linear models*, Bernoulli **19** (2013), no. 4, 1212–1242.

[Bal60] A.V. Balakrishnan, *Estimation and detection theory for multiple stochastic processes*, Journal of Mathematical Analysis and Applications **1** (1960), no. 3, 386–410.

[BCW14] Alexandre Belloni, Victor Chernozhukov, and Lie Wang, *Pivotal estimation via square-root lasso in nonparametric regression*, The Annals of Statistics **42** (2014), no. 2, 757–788.

[Bel02] David A Belsley, *An investigation of an unbiased correction for heteroskedasticity and the effects of misspecifying the skedastic function*, Journal of Economic dynamics and Control **26** (2002), no. 9, 1379–1396.

[BHK+09] Michal Benko, Wolfgang Härdle, Alois Kneip, et al., *Common functional principal components*, The Annals of Statistics **37** (2009), no. 1, 1–34.

[BL08] Peter J Bickel and Elizaveta Levina, *Regularized estimation of large covariance matrices*, The Annals of Statistics (2008), 199–227.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov, *Simultaneous analysis of lasso and dantzig selector*, The Annals of Statistics (2009), 1705–1732.

[BTW+07] Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al., *Sparsity oracle inequalities for the lasso*, Electronic Journal of Statistics **1** (2007), 169–194.

[BVDG11] Peter Bühlmann and Sara Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011.

[CC06] Song Xi Chen and Hengjian Cui, *On bartlett correction of empirical likelihood in the presence of nuisance parameters*, Biometrika **93** (2006), no. 1, 215–220.

[CG14] Song Xi Chen and Bin Guo, *Tests for high dimensional generalized linear models*, arXiv preprint arXiv:1402.4882 (2014).

[CHL03] Song Xi Chen, Wolfgang Härdle, and Ming Li, *An empirical likelihood goodness-of-fit test for time series*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65** (2003), no. 3, 663–678.

[CLS86] P.E. Castro, W.H. Lawton, and E.A. Sylvestre, *Principal modes of variation for processes with continuous sample curves*, Technometrics **28** (1986), no. 4, 329–337.

[CVK09] Song Xi Chen and Ingrid Van Keilegom, *A review on empirical likelihood methods for regression*, Test **18** (2009), no. 3, 415–447.

[CZ10] Song Xi Chen and Ping-Shou Zhong, *Anova for longitudinal data with missing values*, The Annals of Statistics **38** (2010), no. 6, 3630–3659.

[DCL12] Z John Daye, Jinbo Chen, and Hongzhe Li, *High-dimensional heteroscedastic regression with an application to eqtl data analysis*, Biometrics **68** (2012), no. 1, 316–326.

[DHR91] Thomas DiCiccio, Peter Hall, and Joseph Romano, *Empirical likelihood is bartlett-correctable*, The Annals of Statistics **19** (1991), no. 2, 1053–1061.

[Edw84] Anthony William Fairbank Edwards, *Likelihood*, CUP Archive, 1984.

[EH08] R.L. Eubank and Tailen Hsing, *Canonical correlation for stochastic processes*, Stochastic Processes and their Applications **118** (2008), no. 9, 1634–1661.

[Far97] Julian J Faraway, *Regression analysis for a functional response*, Technometrics **39** (1997), no. 3, 254–261.

[FFS10] Jianfeng Feng, Wenjiang Fu, and Fengzhu Sun, *Frontiers in computational and systems biology*, vol. 15, Springer Science & Business Media, 2010.

[FG96] Jianqing Fan and Irene Gijbels, *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66*, vol. 66, Chapman & Hall/CRC, 1996.

[FHL07] Jianqing Fan, Tao Huang, and Runze Li, *Analysis of longitudinal data with semiparametric estimation of covariance function*, Journal of the American Statistical Association **102** (2007), 632–641.

[FL01] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American statistical Association **96** (2001), no. 456, 1348–1360.

[FL08] Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 5, 849–911.

[FZ00] Jianqing Fan and Jin-Ting Zhang, *Two-step estimation of functional linear models with applications to longitudinal data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **62** (2000), no. 2, 303–322.

[GQ65] Stephen M Goldfeld and Richard E Quandt, *Some tests for homoscedasticity*, Journal of the American statistical Association **60** (1965), no. 310, 539–547.

[GVHF11] Jelle J Goeman, Hans C Van Houwelingen, and Livio Finos, *Testing against a high-dimensional alternative in the generalized linear model: asymptotic type i error control*, Biometrika **98** (2011), no. 2, 381–390.

[HM93] Wolfgang Hardle and Enno Mammen, *Comparing nonparametric versus parametric regression fits*, The Annals of Statistics (1993), 1926–1947.

[HMW06] Peter Hall, Hans-Georg Müller, and Jane-Ling Wang, *Properties of principal component methods for functional and longitudinal data analysis*, The annals of statistics (2006), 1493–1517.

[HS81] Harold V Henderson and Shayle R Searle, *On deriving the inverse of a sum of matrices*, Siam Review **23** (1981), no. 1, 53–60.

[HTS+99] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein, *Imputing missing data for gene expression arrays*, 1999.

[JM13] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, arXiv preprint arXiv:1306.3171 (2013).

[KAC+98] Henry K, Erice A, Tierney C, Balfour HH Jr, Fischl MA, Kmack A, Liou SH, Kenton A, Hirsch MS, Phair J, Martinez A, and Kahn JO, *A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced aids. aids clinical trial group 193a study team.*, J Acquir Immune Defic Syndr Hum Retrovirol (1998), 339–349.

[KF00] Keith Knight and Wenjiang Fu, *Asymptotics for lasso-type estimators*, Annals of statistics (2000), 1356–1378.

[KZ13] Seonjin Kim and Zhibiao Zhao, *Unified inference for sparse and dense longitudinal models*, Biometrika (2013), ass050.

[LH08] Peter Langfelder and Steve Horvath, *Wgcna: an r package for weighted correlation network analysis*, BMC bioinformatics **9** (2008), no. 1, 559.

[LH10] Yehua Li and Tailen Hsing, *Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data*, The Annals of Statistics **38** (2010), no. 6, 3321–3351.

[LL14] Weidong Liu and Shan Luo, *Hypothesis testing for high-dimensional regression models*.

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani, *A significance test for the lasso*, Annals of statistics **42** (2014), no. 2, 413.

[LZL+13] Wei Lan, Ping-Shou Zhong, Runze Li, Hansheng Wang, and Chih-Ling Tsai, *Testing a single regression coefficient in high dimensional linear models*.

[Mam93] Enno Mammen, *Bootstrap and wild bootstrap for high dimensional linear models*, The Annals of Statistics (1993), 255–285.

[MB06] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics (2006), 1436–1462.

[MB10] ———, *Stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72** (2010), no. 4, 417–473.

[MC06] Jeffrey S Morris and Raymond J Carroll, *Wavelet-based functional mixed models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68** (2006), no. 2, 179–199.

[MMB09] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann, *P-values for high-dimensional regression*, Journal of the American Statistical Association **104** (2009), no. 488.

[MY09] Nicolai Meinshausen and Bin Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, The Annals of Statistics (2009), 246–270.

[NL14] Yang Ning and Han Liu, *A general theory of hypothesis tests and confidence regions for sparse high dimensional models*, arXiv preprint arXiv:1412.8765 (2014).

[NRWY12] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statist. Sci. **27** (2012), no. 4, 538–557.

[Owe88] Art B Owen, *Empirical likelihood ratio confidence intervals for a single functional*, Biometrika **75** (1988), no. 2, 237–249.

[Owe90] ———, *Empirical likelihood ratio confidence regions*, The Annals of Statistics **18** (1990), no. 1, 90–120.

[Owe01] ———, *Empirical likelihood*, CRC press, 2001.

[PZB+10] Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang, *Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer*, The annals of applied statistics **4** (2010), no. 1, 53.

[QL95] Jin Qin and Jerry Lawless, *Estimating equations, empirical likelihood and constraints on parameters*, Canadian Journal of Statistics **23** (1995), no. 2, 145–159.

[RS91] John A Rice and Bernard W Silverman, *Estimating the mean and covariance structure nonparametrically when the data are curves*, Journal of the Royal Statistical Society. Series B (Methodological) (1991), 233–243.

[Ser80] Robert J Serfling, *Approximation theorems of mathematical statistics*, John Wiley & Sons, 1980.

[SF04] Qing Shen and Julian Faraway, *An f test for linear models with functional responses*, Statistica Sinica **14** (2004), no. 4, 1239–1258.

[Sil78] Bernard W Silverman, *Weak and strong uniform consistency of the kernel estimate of a density and its derivatives*, The Annals of Statistics **6** (1978), no. 1, 177–184.

[SR05] Bernard Walter Silverman and James O. Ramsay, *Functional data analysis*, Springer, 2005.

[SS13] Rajen D Shah and Richard J Samworth, *Variable selection with error control: another look at stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75** (2013), no. 1, 55–80.

[SZ12] Tingni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Biometrika (2012), ass043.

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[TLTT14] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani, *Post-selection adaptive inference for least angle regression and the lasso*, arXiv preprint (2014).

[Tuk77] John W Tukey, *Exploratory data analysis*, Reading, Ma **231** (1977), 32.

[VdG08] Sara A Van de Geer, *High-dimensional generalized linear models and the lasso*, The Annals of Statistics (2008), 614–645.

[vdGBR13] Sara van de Geer, Peter Bühlmann, and Ya'acov Ritov, *On asymptotically optimal confidence regions and tests for high-dimensional models*, arXiv preprint arXiv:1303.0518 (2013).

[Ver10] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027 (2010).

[Wai09] Martin J Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)*, Information Theory, IEEE Transactions on **55** (2009), no. 5, 2183–2202.

[WD12] Jens Wagener and Holger Dette, *Bridge estimators and the adaptive lasso under heteroscedasticity*, Mathematical Methods of Statistics **21** (2012), no. 2, 109–126.

[WR09] Larry Wasserman and Kathryn Roeder, *High dimensional variable selection*, Annals of statistics **37** (2009), no. 5A, 2178.

[WWL12] Lan Wang, Yichao Wu, and Runze Li, *Quantile regression for analyzing heterogeneity in ultra-high dimension*, Journal of the American Statistical Association **107** (2012), no. 497, 214–222.

[XZ07] Liugen Xue and Lixing Zhu, *Empirical likelihood for a varying coefficient model with longitudinal data*, Journal of the American Statistical Association **102** (2007), no. 478, 642–654.

[YMW05a] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang, *Functional data analysis for sparse longitudinal data*, Journal of the American Statistical Association **100** (2005), 577–590.

[YMW05b] ———, *Functional linear regression analysis for longitudinal data*, The Annals of Statistics **33** (2005), no. 6, 2873–2903.

[ZC07] Jin-Ting Zhang and Jianwei Chen, *Statistical inferences for functional data*, The Annals of Statistics **35** (2007), no. 3, 1052–1079.

[Zha09] Tong Zhang, *Some sharp performance bounds for least squares regression with l1 regularization*, The Annals of Statistics **37** (2009), no. 5A, 2109–2144.

[Zha10] Cun-Hui Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics (2010), 894–942.

[Zha11] Jin-Ting Zhang, *Statistical inferences for linear models with functional responses*, Statistica Sinica **21** (2011), no. 3, 1431.

[ZHM+10] Lan Zhou, Jianhua Z Huang, Josue G Martinez, Arnab Maity, Veerabhadran Baladandayuthapani, and Raymond J Carroll, *Reduced rank mixed effects models for spatially correlated hierarchical functional data*, Journal of the American Statistical Association **105** (2010), no. 489, 390–400.

[ZL00] Wenyang Zhang and Sik-Yum Lee, *Variable bandwidth selection in varying-coefficient models*, Journal of Multivariate Analysis **74** (2000), no. 1, 116–134.

[ZPW13] Xiaoke Zhang, Byeong U Park, and Jane-ling Wang, *Time-varying additive models for longitudinal data*, Journal of the American Statistical Association **108** (2013), no. 503, 983–998.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of lasso*, The Journal of Machine Learning Research **7** (2006), 2541–2563.

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 217–242.