Hierarchical Time-varying Mixed-Effects Models in High-Dimensional Time Series and Longitudinal Data Studies

by

Jinglan Li

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

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 12/07/2016

The dissertation is approved by the following members of the Final Oral Committee:

Zhengjun Zhang, Professor, Statistics

Richard Chappell, Professor, Biostatistics & Medical Informatics and Statistics

Mari Palta, Professor, Population Health Sciences

Yazhen Wang, Professor, Statistics

Anru Zhang, Assistant Professor, Statistics

To my family

Acknowledgments

First and foremost, I would like to express my sincere appreciation and gratitude to my advisor Professor Zhengjun Zhang, for the continuous support of my Ph.D. study and related research, for his patience, insightful guidance, and immense knowledge. It has been an honor to be his Ph.D. student. His advice on both research as well as on life have been priceless, whether in academia or in industry or another setting. I immensely learned from every meeting I had with him. His encouragement, motivation, and support are pivotal throughout my Ph.D. life.

Besides, I would like to thank my committee members professor Yazhen Wang, Professor Richard Chappell, Professor Mari Palta, Professor Anru Zhang for their time, stimulating discussions, and brilliant comments on the thesis. I would also like to thank Professor Chunming Zhang for her invaluable feedback on a preliminary version of this thesis.

My sincerest thanks and appreciation also go to the Department of Statistics at the University of Wisconsin-Madison for providing a high-quality education and making my Ph.D. experience great and fruitful. My thanks are also extended to all my academic siblings, Shuang Huang, Qiurong Cui, Bin Zhu, Timothy Idowu, Yaoguo Xie, Ruosi Guo, Lu Yang, Zifeng Zhao, Wenzhi Cao, Yuqing Xu, for their valuable opinions and insightful questions at our weekly group meetings. Additionally, I am truly thankful to my fellow students for their kind support, generous help, and constant encouragement, and all the great memories we shared.

Furthermore, I want to give special thanks to Chao Zhang, for all his love and unconditional support. I would also like to thank all of my friends for always being there to listen and advise.

Last but not least, I want to thank my loving parents, Bin Li and Fangrong Lan. Their love and faith were what sustained me thus far. Words cannot express how grateful I am to my parents for their never-ending encouragement and love, and for always believing in me throughout my life.

Contents

List of Tables

List of Figures

Abstract

Consider a varying coefficient model (Hastie and Tibshirani, 1993), where the coefficient is unknown but is dynamic in the sense that it is a function of a certain covariate. In some cases, the covariate is a special variable 'time'. Motivated by the need for varying-coefficient vector time series models (Jiang, 1999) and varying-coefficient partially linear models (Fan, Huang, and Li, 2007), we are primarily interested in time-varying coefficient models for continuous multivariate time series data and continuous longitudinal data. The challenge is how to simultaneously display serial, clustering, and multivariate attributes of the data set, to which the routinely assumed two-level and univariate response models are not able to apply.

We approach this problem by a flexible new model called multiple response hierarchical time-varying mixed-effects model. So far, the thesis has focused on two responses. Extension to > 2 responses involves no fundamentally new ideas. The model first uses varying-coefficient parameters for accurately describing the dynamic of the series. The new covariance matrix is decomposed into between-response correlation structure of random cluster effect and correlation structure between measurement errors. By allowing shared cluster effects the model allows for characterizing homogeneity in repeated measurements in the same cluster. By allowing for time dependent error terms, it is possible to model the correlation induced by within-subject variation. We adopt a similar approach of Fan and Gijbels (1996), where we first propose local linear regression estimators for the varying coefficients, and then obtain random effect prediction by maximizing the profile likelihood with a closed-form solution. Asymptotic results give good insight into the properties of estimators. It is shown that estimates are consistent. We also conduct the model comparison, it turns out that the proposed methods outperform the traditional univariate response models, nonparametric models, and linear mixed effects models in both predicting the response and estimating the coefficient surface based on simulation studies. Finally, we have applied this model to a real-world study on the price-volume relation of NASDAQ stock market data.

Keywords: Varying coefficient model, Mixed effects model, High-dimensional time series, Multivariate longitudinal data, Local linear regression, Profile likelihood.

Chapter 1

Introduction

The varying coefficient models arise in many situations in economics, finance, political science, social science, and biomedical research areas. They have been successfully applied to generalized linear models, analysis of longitudinal data, functional data, financial and economic data, and in particular time series models as the length of the observed time series increases and the series itself is subject to dynamics of change. These complex data often exhibit nonlinearity, dependence between successive measurements, dependence among individuals within clusters, and positive (negative) dependence between responses. Vast machinery of regression analysis has been developed to the service of modelling this complex data type in the areas of both the time series data and longitudinal data analysis area. Both are concerned with the analysis of data consisting of successive measurements that are recorded over a period of time. For simplicity, we shall regard multiple time series data as a special case of longitudinal data. However, the popular statistical packages (e.g. the R package of Nonlinear Mixed Effect Model) that were being used to combine ideas from varying coefficient models and longitudinal data analysis were revealed as being inadequate when relatively complex covariance matrix structures are used. Furthermore, many models being used to capture the dependence between repeated measurements focused on the subject level (Lin and Carroll, 2000; Lin and Carroll, 2001; Wang, Carroll, and Lin, 2005; Wu and Zhang, 2002; Liu, Ma and O′ Quigley, 2008; Olsen, Delong, Oddone and Bosworth, 2008; Heagerty and Zeger, 2000). It is common that data are clustered within higher levels, e.g., financial sectors, families, or communities. Taking stock market research as an example, it is usually expected that there is dependence between stock prices in the same financial sector and even dependence between certain financial sectors. Poor models of the dependence are thought to be one of the causes of the financial crisis of 2008 (Coval et al., 2009; Zimmer, 2012). Existing cluster-based models, while useful by utilizing the generalized estimating equations, are often built on a working independent correlation matrix, and face the risk of ignoring the dependence structure of the data, i.e., pretending all measurements were independent.

This thesis makes three primary contributions. First, we propose new models for multivariate response longitudinal data based on a hierarchical structure, the use of which makes them particularly attractive for simultaneously modeling serial, clustering, and multivariate attributes of the data set. These models have also been combined with existing nonparametric, semiparametric models to provide flexibility yet tractability in modelling time-varying relationship between the response and covariates. The proposed models allow the researcher to determine the degree of flexibility based on the structure of the data. For example, by allowing for the nonparametric trend function the model features in predicting the trajectory of response over time and by allowing shared cluster effects the model allows for characterizing homogeneity in repeated measures in the same cluster. By allowing for time-dependent error terms, it is possible to capture the dependence within the subjects. Many models made some strong simplifying assumptions on covariance function in order to keep the model tractable (Verbeke and Molenberghs, 2000; Fan, Huang, and Li, 2007; Fan and Wu, 2008), and a prominent feature of the class of proposed models is that such assumptions will not be required and the dependence structure will be employed in the model.

To forecast the trend presented in the longitudinal data, nonparametric models have been considered. Using local polynomial techniques we obtain theoretical results on the multivariate response hierarchical time-varying mixed effects model that we use in our work. Given the data are localized in time with local polynomial techniques, we employ the local linear estimator proposed in Fan, and Gijbels (1996). We also consider an estimation procedure for model coefficients using a profile weighted least squares approach.

The second contribution of this thesis is a study of the estimation of the covariance matrix of multivariate response hierarchical time-varying mixed effects model. This is one of the most complicated covariance matrices of the longitudinal data in the literature. We find significant evidence in favor of a correctly specified within cluster and within subject correlation structure, implying that the common cluster based model which ignores the correlation structure is not suitable for multilevel longitudinal data. Moreover, we find significant evidence that the approach to estimating the covariance function assuming that the correlation structure follows a parametric setting has difficulty in optimization. Empirical results suggest that the estimation procedures commonly implemented using the restricted maximum likelihood (REML) may have difficulty in optimization when the dimension of the cluster effect is bigger than 3 (Verbeke and Molenberghs, 2000).

The third contribution of the thesis is a study of the price-volume relation of 100 stocks of 5 sectors in NASDAQ (National Association of Securities Dealers Automated Quotations) stock market, using daily data over the period of Jan.3 2006 - Dec.31 2015. This is one application of longitudinal data analysis in predicting opening price and trading volume, which is rare in the econometrics literature. Since stocks are clustered within financial sectors, it is natural to assume stocks in the same financial sectors have similar behaviour due to similar economic status. This was particularly prominent when one of the many surprises from the financial crisis of 2008 was the extent to which stock returns that had been previously behaved mostly independent suddenly moved together. In the application to predict the stock market, besides the time-wise correlation within the same stock and the stock correlation within the same sector, a between-response correlation is also considered at each time point.

Certain types of nonparametric and semiparametric models for longitudinal data have already appeared in the literature. The model we consider includes many useful models proposed in the literature. It is a useful extension of commonly used linear models (Diggle et al. ,2002; Demidenko, 2004 and Frees, 2004, for longitudinal data by allowing coefficients to change over time. It is also an extension of useful semiparametric models studied by Fan, Huang and Li (2007) and Sun, Zhang and Tong (2007), in that we retain a semiparametric varying-coefficient partially linear structure, but allow for multiple responses, allow for variables to have shared cluster effect, and relaxes the assumptions on the covariance function. Other work on nonparametric and semiparametric models for longitudinal data are presented in Fan and Zhang (2000), which considers functional linear models for longitudinal data, and in Fan and Li (2004), which discusses parametric inferences and model selection for semiparametric regression models in longitudinal data analysis. Lin and Carrol (2000) and Lin and Carrol (2001) also proposed a nonparametric regression model and a semiparametric generalized linear model, where the entire correlation structure has been ignored. Wu and Zhang (2002) presented a nonparametric mixed-effects model for longitudinal data to estimate both fixed effects curve and random-effects curves. Wang, Carrol and Lin (2005) considered a marginal generalized semiparametric partially linear models, but ignore the within-cluster correlation structure either in nonparametric curve estimation or throughout. With the exception of Fan, Huang and Li (2007) and Fan and Wu (2008), the papers to date have not considered estimation of the correlation structure, instead pretending all observations are independent. Our analysis of multivariate response hierarchical time-varying mixed effects model is new to the literature.

The remainder of the thesis is structured as follows. Chapter 2 presents multiple response hierarchical time-varying mixed effects models for longitudinal data, including its motivations. Followed by Chapter 3 that develops estimation procedures of varying coefficients, random effects covariance matrix, and subject correlation parameters. The consistency of estimators is proved later. Chapter 4 gives a simulation study that compares with alternative models. Chapter 5 presents an empirical study of daily opening price and trading volumes of 100 stocks over the period Jan.3 2006 - Dec.31 2015. Chapter 6 is a discussion of future work. All technical proofs are provided in the last Chapter of Appendix.

Chapter 2

Multiple Response Three-level Time-varying Mixed-effects Models

2.1 Motivation

Vast machinery of regression analysis has been developed to the service of modelling longitudinal data. Many working models can be broadly classified as fully parametric regression methods, fully nonparametric regression methods, and semi-parametric regression methods.

There is a huge body of literature on parametric models for longitudinal data analysis, including linear mixed-effects models (see, e.g., Laird and Ware, 1982; Diggle et al., 2002 and the references therein). Parametric mean models enjoy simplicity and their properties are very well established, they are typically limited by the inflexibility in modeling complicated relationships between the response and covariates in various longitudinal studies since they assume that the mean of response to covariates is purely parametric which may not always hold in applications. Furthermore, parametric models involving a finite set of parameters do not work well when there is a large number of repeated measurements, as the amount of information about the data may not be fully captured by only a few parameters.

To relax the assumptions on tightly specified parametric models in describing the relationship between a longitudinal response and covariates, various nonparametric models have been proposed, see Fan and Zhang (2000); Lin and Ying (2001) and references therein for theory and applications of nonparametric models. Although fully nonparametric approaches are appealing and make no assumptions on the specification of the model, they are hard to interpret and incapable of incorporating some prior information for the unknown functions. Worse still, they may suffer from the so-called ’curse of dimensionality’ problem, which rules out the standard nonparametric methods in modern applications for the increasing high-dimensional data sets. To ease the ’curse of dimensionality’, many powerful dimensionality reduction approaches are proposed. Examples include additive models, see Friedman and Stuetzle (1981), partially linear models, see Engle, Granger, Rice, and Weiss (1986); and varying coefficient models, see Cleveland, Grosse and Shyu (1991); Hastie and Tibshirani (1993).

Semiparametric regression modeling has been used extensively in the literature because it provides the flexibility of fully nonparametric approaches in modeling complexity of the data sets while maintaining the model interpretability of fully parametric methods. Ruppert, Wand, and Carroll (2009) summarized applications and theoretical developments of semiparametric regression models. The partially linear model, the most commonly used semiparametric regression models, is gaining a lot of attention in literature in recent years. Engle, Granger, Rice, and Weiss (1986) introduced the partially linear model defined by

$$Y_i = g(X_i) + Z_i^T \beta + \epsilon_i, \tag{2.1}$$

where $X_i = (X_{i1}, \ldots, X_{id})^T$ and $Z_i = (Z_{i1}, \ldots, Z_{ip})^T$ are vectors of explanatory variables, (X_i, Z_i) are considered to be i.i.d, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters, g is an unknown function from $\mathbb{R}^d$ to $\mathbb{R}^1$, and $\epsilon_1, \ldots, \epsilon_n$ are independent random errors with mean zero and finite variances σ_i^2. Partially linear models have many applications, particularly in economics, finance, and biology. Engle, Granger, Rice and Weiss (1986) were among the first to analyze the relationship between temperature and electricity usages using these models.

On the other hand, several authors proposed the varying coefficient models. There are two ways in building a varying coefficient model. Hastie and Tibshirani (1993) introduced the varying coefficient models, formulated as follows:

$$Y_i = \alpha_0 + X_{i1}\alpha_1(U_{i1}) + \cdots + X_{ip}\alpha_p(U_{ip}) + \epsilon_i, \tag{2.2}$$

where $\alpha_1, \ldots, \alpha_p$ are p separate unknown varying coefficient functions that needs to be estimated, $X_1, \ldots, X_p$ and $U_1, \ldots, U_p$ are covariates and ϵ_i is random error with mean 0 and finite variances σ_i^2. Model (2.2) lets the coefficients of the $X_1, \ldots, X_p$ be functions of different covariates $U_1, \ldots, U_p$, which change the effects of the covariates $X_1, \ldots, X_p$ in nonparametric ways. A second class of models is that where all regression coefficients $\alpha_1, \ldots, \alpha_p$ are assumed to depend on a single covariate U, see Fan and Zhang (1999), which leads to the model

$$Y_i = \alpha_0 + X_{i1}\alpha_1(U_i) + \cdots + X_{ip}\alpha_p(U_i) + \epsilon_i, \tag{2.3}$$

for given response variable Y and covariates $X_1, \ldots, X_p, U$ with the unknown varying coefficient functions $\alpha_1, \ldots, \alpha_p$. In this thesis, our key interest is model (2.3). Like the partially linear models, the varying coefficient models is one way of freeing us from the curse of dimensionality and reducing the modeling bias. Another advantage of the varying coefficient models is that they are easy to interpret. And based on the assumption that the regression coefficients are smooth functions of the covariates, they allow us to examine the covariate effects vary over different groups characterized by some covariates such as time. When the covariate is time, we call them time-varying coefficient models.

Various extensions of models (2.1) and (2.3) have been widely adopted in the literature, including those that accommodate longitudinal responses, discrete responses and in particular time series as the length of the observed time series increases and the series itself is subject to dynamics of change. Hoover et al. (1998) and Fan and Zhang (2000) extended the varying coefficient model (2.3) for Gaussian longitudinal outcomes by allowing the error terms to be correlated within the same subject. Suppose that Y_{ij} and T_{ij} are the outcome and the covariate of the jth observation from the ith subject, and

$$Y_{ij} = X_{ij}^T(T_{ij})\alpha(T_{ij}) + \epsilon_{ij}, \tag{2.4}$$

where $X_i(t)$ is a $p \times 1$ vector of covariates and $\alpha(t) = (\alpha_0(t), ..., \alpha_p(t))^T$ is a $p \times 1$ vector of unspecified smooth functions. Zhang (2004) proposed generalized linear mixed models with varying coefficients for discrete longitudinal data. Jagannathan and Wang (1996), Reyes (1999), Akdeniz et al. (2003), and Cai (2007) studied time-varying coefficient time series models. In case the response vector is subject to changes over time, we can replace y_i by a vector, say the vector time series, where coefficients are coefficient matrices for $i = 1, \ldots, n$. The resulting model is in Jiang (1999), which used the time-varying coefficient vector autoregressive (VAR) model for multivariate time series, defined as

$$y_i = A_0(t_i) + A_1(t_i)y_{i-1} + \cdots + A_p(t_i)y_{i-p} + \epsilon_i, \tag{2.5}$$

where y_i is a vector, $A_0(t), \ldots, A_p(t)$ are matrices varying with t, y_{i-1} is called the l-th lag of y, and ϵ_i is a vector of error terms with mean 0 and covariance matrix Σ. Applications of time-varying coefficient models in time series include the return of a market index effect on the return of an individual stock, where the coefficient is called beta-coefficient in the capital asset pricing model (CAPM) (see, Cochrane, 2001 and Tsay, 2002). Studies show that the beta-coefficient might change over time (Cai, 2007); and the demand elasticity in economic-causal model, see for example Hackl and Westlund (1995, 1996); Orbeetal (2006). In the price-volume relation study described in Chapter 5, 100 NASDAQ stocks were examined every day for up to ten consecutive years for the dynamic properties of long time series.

The introduction of varying coefficients in partially linear models leads to an important extension of the partially linear model, the varying coefficient partially linear model, which are frequently used in practice. Let Y be the response variable and $\{U, X, Z\}$ be its covariates. The varying-coefficient partially linear model (Fan and Huang, 2005) is defined to be

$$Y_i = X_i^T \alpha(U_i) + Z_i^T \beta + \epsilon_i, \tag{2.6}$$

In this thesis, we extend the varying coefficient partially linear model (2.6) to consider multivariate longitudinal response with a three-level hierarchical structure. Allowing for more than a single outcome improves the efficiency of the estimated coefficients because the correlation between outcomes is able to be considered. The researchers, on the other hand, are usually able to draw on the large literature on models for a single response. The extensions to consider the case that the number of responses is greater than two is not analyzed here since the generalization of three or more responses is not difficult.

Multilevel models in extant literature generally have two levels, where in the first level, correlated measurements are recorded, and in the second level, several independent clusters of correlated measurements are observed. Lin and Carroll (2001) built a two-level hierarchical model. In their model, the repeated measurements of the presence of respiratory infection on preschool-age children are correlated, and serve as the first level. These measurements are clustered within the same children, which serves as the second level. Though widely applied, a two-level model may not be suitable for all situations. For example, stock prices are recorded for each company in the stock market every day. In this case, a two-level model is not suitable, because stock companies are grouped in different sections according to their business nature. So a three-level model would be more appropriate in this situation.

Lin and Carroll (2006) proposed a varying-coefficient three-level model that allows between-subject and within-subject correlations. They consider the problem in which there are n families, family i has L_i children. For the jth child, $j = 1, \ldots, L_i$, the repeated measurements y_{ijk}, a base-line measure Z^*_{ij}, covariates X_{ijk} are collected over time for $k = 1, \ldots, m_{ij}$. The model is:

$$y_{ijk} = X^T_{ijk}\beta_0 + \theta_0(Z^*_{ij}) + \epsilon_{ijk}, \tag{2.7}$$

where ϵ_i has a covariance matrix Σ_i, which is $\sum_{j=1}^{L_i} m_{ij} \times \sum_{j=1}^{L_i} m_{ij}$ matrix. Σ_i allows them to model the dependencies within repeated measurements and subjects, but the estimation of Σ_i is not fully developed in the paper. Later in the simulation part of the paper, the covariance matrix is estimated as the sample covariance matrix of the residuals, which is very rough. They further assumed that the correlation structure is autoregressive between repeated measures over time and common between-subject correlation within the same family. Under this setting, the correlation between one child's first and last measurements may be smaller than the correlation between one child's first measurement and another child's last measurement. This motivates a new model and covariance structure. In next section, we present a multivariate three-level time-varying mixed-effects model by incorporating varying coefficients, multivariate outcomes and clusters into (2.7).

2.2 Model Specification

Suppose $(y_{1ijk}, \ldots, y_{dijk})^T$ is a $d \times 1$ vector of response variables taken on subject j in the ith cluster at time points t_{ijk}. Assume without loss of generality that, $d = 2$, $(y_{1ijk}, y_{2ijk})^T$. Consider the model

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} x_{ijk1}\alpha_{11}(t_{ijk}) + \cdots + x_{ijkp}\alpha_{1p}(t_{ijk}) \\ x_{ijk1}\alpha_{21}(t_{jik}) + \cdots + x_{ijkp}\alpha_{2p}(t_{ijk}) \end{pmatrix} + \begin{pmatrix} z_{ijk1}\beta_{1i1} + \cdots + z_{ijkq}\beta_{1iq} \\ z_{ijk1}\beta_{2i1} + \cdots + z_{ijkq}\beta_{2iq} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix}, \tag{2.8}$$

$i = 1, \ldots, n_c$, $j = 1, \ldots, n_s$, and $k = 1, \ldots, n_m$, where n_c is the number of clusters, n_s is the number of subjects, n_m is the number of balanced measurements; $\alpha_{11}(t_{ijk})$,...,$\alpha_{1p}(t_{ijk})$, $\alpha_{21}(t_{ijk})$,...,$\alpha_{2p}(t_{ijk})$ are unspecified smooth functions, β_{1i1},..., β_{1iq}, β_{2i1},..., β_{2iq} are random effects, and ϵ_{dijk} are mean-zero error terms that are uncorrelated with other covariates. The model allows two responses y_{1ijk}, y_{2ijk}. Since two responses are modeled together, the correlation between corresponding random effects should be taken into account. With $\beta_{di} = (\beta_{di1}, \ldots, \beta_{diq})^T$, and $\beta_i = (\beta_{1i}^T, \beta_{2i}^T)^T$, we further assume that $\{\beta_i, i = 1, \ldots, n_c\}$ are i.i.d. $N(\mathbf{0}, D)$, where D is $2q \times 2q$ covariance matrix of β_i. Let us rewrite the model in (2.8) briefly as

$$y_{ijk} = \begin{pmatrix} \alpha_1(t_{ijk})^T \\ \alpha_2(t_{ijk})^T \end{pmatrix} x_{ijk} + \begin{pmatrix} \beta_{1i}^T \\ \beta_{2i}^T \end{pmatrix} z_{ijk} + \epsilon_{ijk}. \tag{2.9}$$

In (2.9), $y_{ijk} = (y_{1ijk}, y_{2ijk})^T$ is a 2×1 vector of response variables, $\alpha_d(t_{ijk}) = (\alpha_{d1}(t_{ijk}), \ldots, \alpha_{dp}(t_{ijk}))^T$ is a $p \times 1$ vector of unknown smooth functions, $d = 1, 2$; $\beta_{di} = (\beta_{di1}, \ldots, \beta_{diq})^T$ is a $q \times 1$ vector of random effects across the clusters; $\beta_i = (\beta_{1i}^T, \beta_{2i}^T)^T$ are i.i.d. $N(\mathbf{0}, D)$, where D is $2q \times 2q$ covariance matrix of β_i; $x_{ijk} = (x_{ijk1}, \ldots, x_{ijkp})^T$ ($p \times 1$) and $z_{ijk} = (z_{ijk1}, \ldots, z_{ijkp})^T$ ($q \times 1$) are covariates, that are modeled as realizations of some stochastic processes; $\epsilon_{ijk} = (\epsilon_{1ijk}, \epsilon_{2ijk})^T$ is a 2×1 vector error term with $E(\epsilon_{ijk}|x_{ijk}, z_{ijk}) = \mathbf{0}$, which implies that x_{ijk} and z_{ijk} are uncorrelated with ϵ_{ijk}. Note that x_{ijk} and z_{ijk} are allowed to be stationary or nonstationary.

The error term in model (2.8) is assumed to follow a mean-reverting Ornstein-Uhlenbeck process (Doob, 1942) with mean zero. The time points are regarded as a random sample according to a probability density function, namely, $f(t)$. In some cases, t_{ijk} might be equally spaced over the interval of interest. We denote the support of $f(t)$ or the interval of interest as $\mathcal{T}$, which is a finite interval. To facilitate the presentation, we assume that $\mathcal{T}$ is [0,T] and $\epsilon_{dijk} = \epsilon_{dij}(t_k), k = 1, \ldots, n_m$, which is an Ornstein-Uhlenbeck process. To be more specific, $\epsilon_{dij}(t_k)$ satisfies the following stochastic differential equation

$$\begin{cases} d\epsilon_{dij}(t) = -\theta_d \epsilon_{dij}(t) dt + \sigma_d dW(t) \\ \epsilon_{dij}(0) \sim N(0, \frac{\sigma_d^2}{2\theta_d}), \end{cases} \tag{2.10}$$

where $\theta_d > 0$ and $\sigma_d > 0$ are parameters, W_t is the Wiener process and $\epsilon_{dij}(0)$ is independent of the Wiener proces $W(t)$. The solution to (2.10) is

$$\epsilon_{dij}(t) = \epsilon_{dij}(0)e^{-\theta_d t} + \sigma_d \int_0^t e^{-\theta_d (t-s)}\, dW(s). \tag{2.11}$$

Three fundamental properties of the Ornstein-Uhlenbeck process (Ricciardi and Sacerdote, 1979) which will be useful are:

1. The Ornstein-Uhlenbeck process is stationary.

2. The Ornstein-Uhlenbeck process is Markov.

3. For $s < t$, $\epsilon_{dij}(t)$ and $\epsilon_{dij}(s)$ have a bivariate Gaussian distribution. Furthermore, $\epsilon_{dij}(t)$ have a Gaussian distribution with mean $E\epsilon_{dij}(t) = 0$ and variance $\mathrm{Var}(\epsilon_{dij}(t)) = \frac{\sigma_d^2}{2\theta_d}$, and correlation coefficients determined by the equation

$$\mathrm{Cov}(\epsilon_{dij}(t), \epsilon_{dij}(s)) = \frac{\sigma_d^2}{2\theta_d} e^{-\theta_d |t-s|}. \tag{2.12}$$

From (2.12), we see immediately that the correlation coefficients of the process decrease exponentially, and the Ornstein-Uhlenbeck process imposes an AR(1) correlation structure on the time-wise correlation between error terms. Besides, the Ornstein-Uhlenbeck process is widely used to model continuous-time error terms. For example, in an analysis of continuous-time change in multivariate longitudinal data, Oravecz, Tuerlinckx, and Vandekerckhove(2009) applied the Ornstein-Uhlenbeck process to model the correlation of the observations from a single person; Wang, and Taylor (2001) used the Ornstein-Uhlenbeck processes to account for the random fluctuations around the population average. In finance, the Ornstein-Uhlenbeck process is used to model the departure of instantaneous interest rate from the long term mean level in Vasicek model, see Hull (2006) .

In the next chapter, we will study the local polynomial techniques and the profile likelihood, thereby explore estimation methods starting from the estimation of the varying coefficient functions.

Chapter 3

Local Polynomial Method and Profile-likelihood Approach

There are many approaches to estimate the varying coefficient functions $\alpha_d(t_{ijk}) = (\alpha_{11}(t_{ijk})^T, \ldots, \alpha_{1p}(t_{ijk}))$ and predict the random effects $\beta_{di} = (\beta_{di1}, \ldots, \beta_{diq})^T$. The profile least squares is a useful approach and will be shown to be consistent for model (2.9).

In this chapter, we will start with a review of the local polynomial techniques, then move to the method of profiling, followed by the profile least squares approach, which leads us to estimators for unknown functions $\alpha_d(.)$ and predictors for β_{di} in the multivariate response three-level time-varying mixed-effects model. This is followed by hypothesis testing.

3.1 Local Polynomial Techniques

3.1.1 Introduction to Local Polynomial Regression

Nonparametric and semiparametric regression motheds for longitudinal data using kernel and spline methods have been well developed during the past years, e.g., Zeger and Diggl (1994), Hoover et al. (1998), Fan and Zhang (2000), Lin and Ying (2001), Carroll and Lin (2004), and Fang and Li (2004). Consider the following simple nonparametric regression model

$$Y_i = g(T_i) + \epsilon_i, \tag{3.1}$$

where $t_i, i = 1, \dots, n$ are time points, y_i are the responses at time points, g is an unknown function, and $\epsilon_i \sim N(0, \sigma^2)$ and are i.i.d ($i = 1, \dots, n$). Most important, in nonparametric regression we do not assume that g has a certain parametric form. For longitudinal data, Model (3.1) is for each subject, where g is the individual function, and $t_i, i = 1, \dots, n$ are the individual time points of n repeated measurements.

There are many existing smoothing methods that can be used to estimate the g in (3.1). Local polynomial regression, the most popular kernel regression method, has attractive properties over other smoothing methods such as the classical Nadaray-Watson kernel method both from a theoretical and practical point of view. See Wand and Jones (1994); Fan and Gijbels (1996) for some discussions on several major advantages of using the local polynomial regression. We are going to briefly introduce the idea of the local polynomial regression in this subsection.

The main idea of local polynomial regression is to locally approximate the g in

(3.1) using data around any arbitrary point t by a polynomial of some degree. Taylor series expansion is the foundation of this method, which states that every smooth function can be approximated locally by polynomials if the function is smooth.

To be more specific, let t be any fixed time point where the g will be estimated. By assuming that $g(.)$ has a $p+1$ continuous derivative, $g(.)$ can be approximated by a polynomial of degree p as $g(t_i) \approx g(t) + g\prime(t)(t_i - t) + \ldots + g^{(p)}(t)(t_i - t)^p/p! =: \sum_{j=0}^{p} a_j (t_i - t)^j$ in a neighborhood of t, where $g^{(p)}(t)$ denotes the p-th derivative of g at t, and

$$a_j = g^{(j)}(t)/j!, j = 0, \ldots, p. \tag{3.2}$$

Typically, it is of interest to estimate $g(t)$. Let $\hat{a}_j$, $j = 0, \ldots, p$ be the minimizer of the following negative local log-likelihood, apart from a constant, defined as

$$\frac{1}{2\sigma^2} \sum_{i=1}^{n} K_h(t_i - t) \left\{ y_i - \sum_{j=0}^{p} a_j (t_i - t)^j \right\}^2, \tag{3.3}$$

where $K_h(s) = K(s/h)/h$, a rescaling of a kernel function $K(.)$ with a constant $h > 0$. h is called bandwidth or smoothing parameter. It controls the degree of smoothing and specifies the size of the local neighborhood around t, namely, $I_h(t) = [t-h, t+h]$, which allows the local fitting. $K(.)$ is often taken to be a probability density. And most nonparametric estimation uses symmetric kernels, such as Gaussian, uniform, and Epanechnikov kernels.

Then according to (3.2), $g^{(j)}(t) = j! a_j$, $j = 0, \ldots, p$. In particular, the resulting local polynomial estimator of $g(t)$ is $\hat{g}(t) = \hat{a}_0$.

Local constant kernel estimator and local linear kernel estimator are two simplest estimators. The local constant kernel estimator is known as the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), which results from (3.3) by simply taking $p = 0$,

$$\hat{g}(t) = \frac{\sum_i^n K_h(t_i - t)y_i}{\sum_1^n K_h(t_i - t)}. \tag{3.4}$$

Within a local neighborhood, it approximates the regression function by a (local) constant. When the kernel function is the Uniform kernel, the Nadaraya-Watson estimator is exactly the local average of y_i's. The local linear estimator ($p = 1$) is obtained by approximating the function locally with a linear function, see Fan (1992). Let $(\hat{a}_0, \hat{a}_1)$ minimize the following

$$\frac{\sum_i^n K_h(t_i - t)[y_i - a_0 - a_1(t_i - t)]}{\sum_1^n K_h(t_i - t)}, \tag{3.5}$$

then the local linear estimator is $\hat{g}(t) = \hat{a}_0$, which has the expression as

$$\hat{g}(t) = \frac{\sum_{i=1}^n [s_2(t) - s_1(t)(t_i - t)]K_h(t_i - t)y_i}{s_2(t)s_0(t) - s_1^2(t)}, \tag{3.6}$$

where $s_r(t) = \sum_{i=1}^n K_h(t_i - t)(t_i - t)^r$, $r = 0, 1, 2$. The local linear estimator has been commonly used because of its better properties at the boundary. See discussions on these properties in Hastie and Loader (1993), and Fan and Gijbels (1996).

3.1.2 Local Polynomial Estimator for the Multivariate Response Three-level Time-varying Mixed-effects Model

We have proposed the multivariate response three-level time-varying mixed-effects model (2.9) in section 2.2, as

$$y_{ijk} = \begin{pmatrix} \alpha_1(t_{ijk})^T \\ \alpha_2(t_{ijk})^T \end{pmatrix} x_{ijk} + \begin{pmatrix} \beta_{1i}^T \\ \beta_{2i}^T \end{pmatrix} z_{ijk} + \epsilon_{ijk},$$

where $y_{ijk} = (y_{1ijk}, y_{2ijk})^T$ is a 2×1 vector of response variables, $\alpha_d(t_{ijk}) = \{\alpha_{d1}(t_{ijk}), \dots, \alpha_{dp}(t_{ijk})\}^T$ is a $p \times 1$ vector of unknown smooth functions, $d = 1, 2$; $\beta_{di} = \{\beta_{di1}, \dots, \beta_{diq}\}^T$ is a $q \times 1$ vector of random effects across the clusters; $\beta_i = (\beta_{1i}^T, \beta_{2i}^T)^T$ are i.i.d. $N(\mathbf{0}, D)$, where D is $2q \times 2q$ covariance matrix of β_i; $x_{ijk} = (x_{ijk1}, \dots, x_{ijkp})^T$ $(p \times 1)$ and $z_{ijk} = (z_{ijk1}, \dots, z_{ijkp})^T$ $(q \times 1)$ are covariates; $\epsilon_{ijk} = (\epsilon_{1ijk}, \epsilon_{2ijk})^T$ is a 2×1 vector error term. Hoover et al. (1998) first introduced and discussed the local polynomial regression method for time-varying coefficient models for longitudinal data. Fan and Zhang (2000), Huang et al.(2002) further developed the estimation procedures among others. Fang, Huang and Li (2007) showed that the varying coefficient, estimated by the local polynomial regression in the semiparametric models, does not greatly affected by ignoring the covariance structure since the data are localized in time.

We adopt a similar approach of Fan, Huang and Li (2007), and obtain the profile least squares estimators of the varying coefficients and the random effects, which have a closed form. To get the profile least squares estimators, we first consider the local linear estimator, i.e., a polynomial with degree 1, for $\alpha_d(.)$, $d = 1, 2$, constructed by ignoring the within-subject and within-cluster correlation. Note that local fitting of polynomials has been applied separately to each response. Assume that $\alpha_d(.)(p \times 1)$ has a second continuous componentwise deriatives. For each given time point t, $\alpha_d(t_{ijk})$ can be componentwisely approximated by a linear function for t_{ijk} in a neighborhood of t, that is,

$$\alpha_d(t_{ijk}) \approx a_{d0} + a_{d1}(t_{ijk} - t),$$

where $a_{d0} = (a_{d0,1}, \ldots, a_{d0,p})^T$, $a_{d1} = (a_{d1,1}, \ldots, a_{d1,p})^T$, and p is the dimension of $\alpha_d(t_{ijk})$. The negative local log-likelihood is

$$\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \Big(\tilde{y}_{dijk} - x_{ijk}^T\big(a_{d0} + a_{d1}(t_{ijk} - t)\big)\Big)^2 K_{h_d}(t_{ijk} - t), \tag{3.7}$$

where $\tilde{y}_{dijk} = y_{dijk} - z_{ijk}^T \beta_{di}$, $h_d > 0$ is the bandwidth and $K_h(t) = K(t/h)/h$, where $K(.)$ is the kernel function. Minimizing the negative local log-likelihood (3.7) gives estimate $(\hat{a}_{d0}, \hat{a}_{d1})$, the components in $\hat{a}_{d0}$ gives estimates of $a_{d0,1}, \ldots, a_{d0,p}$. The estimators obtained from the caculations are

$$\begin{aligned}(\hat{a}_{d0}^T, \hat{a}_{d1}^T)^T = &\Bigg[\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_{ijk} - t)\begin{pmatrix} x_{ijk} \\ (t_{ijk} - t)x_{ijk}\end{pmatrix}(x_{ijk}^T, (t_{ijk} - t)x_{ijk}^T)\Bigg]^{-1} \\ &\times \Bigg[\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_{ijk} - t)\begin{pmatrix} x_{ijk} \\ (t_{ijk} - t)x_{ijk}\end{pmatrix}\tilde{y}_{dijk}\Bigg].\end{aligned} \tag{3.8}$$

Then estimator of $\alpha_d(t, \beta)$ is simply $\hat{a}_{d0}(p \times 1)$, which is inferred from (3.8),

$$\hat{\alpha}_d(t, \beta) = [s_0(t) - s_1(t)s_2(t)^{-1}s_1(t)]^{-1} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \frac{[I_p - s_1(t)s_2(t)^{-1}(t_{ijk} - t)]K_h(t_{ijk} - t)x_{ijk}\tilde{y}_{dijk}}{n_c n_s n_m}, \tag{3.9}$$

where I_p is a $p \times p$ identity matrix and $s_r(t) = \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_{ijk} - t)(t_{ijk} - t)^r x_{ijk} x_{ijk}^T$, $r = 0, 1, 2$. Then, the local linear estimator of $\alpha_d(t)$ is $\hat{\alpha}_d(t, \hat{\beta})$. Note that, $s_r(t), r = 0, 1, 2$, is a $p \times p$ matrix. It is easily seen that the local linear estimator is a local weighted average, where data points, around the targeted time point t, have different weight loadings according to the kernel function $K(.)$.

3.2 The Method of Profiling

In this section, we focus on the method of profiling and its extension to the semiparametric models. Applying the estimation method described, we propose predictors for β in section 3.3.

3.2.1 The Profile Likelihood

Suppose that we have a random sample of n observations, $y_1, \ldots, y_n$, from a distribution depending on unknown parameters θ, which can be partitioned as $\theta = (\alpha, \beta)$, here $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ are the $p \times 1$ parameters and $\beta = (\beta_1, \ldots, \beta_q)^T$ are the $q \times 1$ parameters. We will need to estimate both α and β.

The log-likelihood is $l(\alpha, \beta) = \sum_{i=1}^{n} l_i(\alpha, \beta)$, where $l_i(\alpha, \beta)$ is the log-likelihood for y_i. To estimate α and β, one can find $(\hat{\alpha}, \hat{\beta}) = argmax_{\alpha,\beta} l(\alpha, \beta)$. However, this can be quite difficult, and leads to expression which is hard to maximize. Instead we consider a different method, which may sometimes be easier to evaluate. The profile likelihood approach is as follows. The log profile likelihood function for β is

$$l(\beta) = l\big(\hat{\alpha}(\beta), \beta\big), \tag{3.10}$$

where $\hat{\alpha}(\beta)$ is the maximum likelihood estimate of α (for fixed β). Importantly, we have 'profiled out' α, and the likelihood is completely in terms of β. The solution to (3.10) is $\hat{\beta} = argmax_{\beta} l(\beta) = argmax_{\beta} l\big(\hat{\alpha}(\beta), \beta\big)$. It is clear that the maximum profile likelihood estimate $\hat{\beta}$ is the same as the overall maximum likelihood estimate.

3.2.2 Extension to the Semiparametric Cases

Now let us consider the extension of the above results to the case when the nuisance parameters α are fully nonparametric functions $\alpha_i(.), i = 1, \ldots, p$ and hence the parameter space is infinite-dimensional.

Speckman (1988) introduced the idea of profile least-squares for the partially linear model. Profile least squares is a useful approach in many semiparametric problems. When $y_i \sim N(0, \sigma^2)$, the approach becomes profile likelihood estimation, see Fan and Huang (2005). Severini and Wong (1992) studied the profile likelihood estimation for the semiparametric class of models, which is shown to be an asymptotically efficient estimator. In the next section, we will integrate the idea of the profile likelihood into the multivariate response three-level time-varying mixed-effects model.

3.3 Prediction of Random effects

Suppose that $y_{dijk}, d = 1, 2, i = 1, \ldots, n_c, j = 1, \ldots, n_s, k = 1, \ldots, n_m$, are a random sample from model (2.9). For the jth subject in the ith cluster, $i = 1, \ldots, n_c$, $j = 1, \ldots, n_s$, assume that the covariates x_{ij}, z_{ij} are stochastic processes collected at time points $t_k, k = 1, \ldots, n_m,$, such that $x_{ijk} = x_{ij}(t_k)$, $z_{ijk} = z_{ij}(t_k)$. Under this assumption, $\{x_{ijk}, k = 1, \ldots, n_m\}$ could be independent, dependent, or constant, which is more realistic for real problems. The estimation for β and α can be based on the profile likelihood method.

The profile likelihood estimator of (α, β) has a closed form using the following matrix notation. In what follows, we propose to adopt the matrix notations.

$$
\begin{array}{lll}
Y_{dij} = (y_{dij1}, \ldots, y_{dijn_m})^T, & Y_{di} = (Y_{di1}^T, \ldots, Y_{din_s}^T)^T, & Y_d = (Y_{d1}^T, \ldots, Y_{dn_c}^T)^T; \\
X_{ij} = (x_{ij1}, \ldots, x_{ijn_m})^T, & X_i = (X_{i1}^T, \ldots, X_{in_s}^T)^T, & X = (X_1^T, \ldots, X_{n_c}^T)^T; \\
Z_{ij} = (z_{ij1}, \ldots, z_{ijn_m})^T, & Z_i = (Z_{i1}^T, \ldots, Z_{in_s}^T)^T, & Z = \mathrm{diag}(Z_1, \ldots, Z_{n_c}); \\
E_{dij} = (\epsilon_{dij1}, \ldots, \epsilon_{dijn_m})^T, & E_{di} = (E_{di1}^T, \ldots, E_{din_s}^T)^T, & E_d = (E_{d1}^T, \ldots, E_{dn_c}^T)^T; \\
m_{dij} = (x_{ij1}^T \alpha_d(t_1), \ldots, x_{ijm}^T \alpha_d(t_m))^T, & m_{di} = (m_{di1}^T, \ldots, m_{din_s}^T)^T, & m_d = (m_{d1}^T, \ldots, m_{dn_c}^T)^T; \\
\beta_d = (\beta_{d1}^T, \ldots, \beta_{dn_c}^T)^T, & \beta = (\beta_1^T, \beta_2^T)^T. &
\end{array}
\tag{3.11}
$$

Then the model (2.9) can be further written as

$$
\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} + \begin{pmatrix} Z & \\ & Z \end{pmatrix} \beta + \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}. \tag{3.12}
$$

For known G, and R_d, $d = 1, 2$, where $G = \mathrm{Var}(\beta)$, and $R_d = \mathrm{Var}\, E_d$, the profile likelihood estimate of β can be obtained from minimizing the following negative log-likelihood of the joint density function of $y_{dijk}, d = 1,2, i = 1,\dots,n_c, j = 1,\dots,n_s, k = 1,\dots,n_m$, and $\beta_i, i = 1,\dots,n_c$, (up to a constant),

$$\begin{aligned} l(\beta) =& \beta^T G^{-1} \beta \\ &+ \log|R_1| + \log|R_2| + \log|G| \\ &+ (Y_1 - m_1 - Z_1\beta_1)^T R_1^{-1} (Y_1 - m_1 - Z_1\beta_1) + (Y_2 - m_2 - Z_2\beta_2)^T R_2^{-1} (Y_2 - m_2 - Z_2\beta_2), \end{aligned} \tag{3.13}$$

where $Y_1, Y_2, m_1, m_2, Z_1, Z_2, \beta_1, \beta_2$ and β are defined in (3.11). Since $\beta_i, i = 1,\dots,n_c$, are random-effects parameter vectors, the expression (3.13) is not a conventional log-likelihood. Note that the first term of the right-hand side of (3.13) is a penalty due to random-effects β taking the between-cluster variation into account, and the last two terms are a weighted residual taking the within-subject variation into account.

From (3.9), we can see that the local linear estimator for $\alpha_d(t)$ results in a linear estimate in $\tilde{y}_{dijk}$, thus $\hat{\alpha}_d(t)$ is linear in $y_{dijk} - z_{ijk}^T \beta_{di}$. The estimate of m_{dij} is $\hat{m}_{dij} = (x_{ij1}^T \hat{\alpha}_d(t_1), \dots, x_{ijm}^T \hat{\alpha}_d(t_m))$, then obviously $\hat{m}_d$ can be expressed as a linear transformation of $Y_d - Z\beta$, that is, there is a $n_c n_s n_m \times n_c n_s n_m$ matrix S_d such that $\hat{m}_d = S(Y_d - Z\beta)$. The matrix S_d is usually called a smoothing matrix. Indeed, we can derive an explicit form of S. Substituting $\hat{m}_d = S_d(Y_d - Z\beta)$ into equation (3.13), for known G, and R_d, $d = 1,2$, the expression (3.13) is minimized by

$$\hat{\beta} = V^{-1} \begin{pmatrix} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Y_1 \\ Z^T (I - S_2)^T R_2^{-1} (I - S_2) Y_2 \end{pmatrix} \tag{3.14}$$

where

$$V = \begin{pmatrix} Z^T(I-S_1)^T R_1^{-1}(I-S_1)Z & \\ & Z^T(I-S_2)^T R_2^{-1}(I-S_2)Z \end{pmatrix} + G^{-1}$$

Thereby, the prediction of β is explicit, i.e., non-iterative. The profile likelihood estimator for the varying coefficient function α_d is simply

$$\hat{\alpha}_d(t) = \hat{\alpha}_d(t, \hat{\beta}). \tag{3.15}$$

It may be nature to consider estimating α and β through iterations. More specific, for given m, the solution $\hat{\beta}$ minimizing the expression (3.13) is

$$\hat{\beta} = \left[\begin{pmatrix} Z^T R_1^{-1} Z & \\ & Z^T R_2^{-1} Z \end{pmatrix} + G^{-1} \right]^{-1} \begin{pmatrix} Z^T R_1^{-1}(Y_1 - m_2) \\ Z^T R_2^{-1}(Y_1 - m_2) \end{pmatrix}.$$

Starting from an initial estimation $\hat{\alpha}$, calculate $\hat{m}$, and next according to $\hat{m}$, estimate β; then go back and use $\hat{\beta}$ to update $\hat{\alpha}$, which can in turn be used to update $\hat{\beta}$. This estimation procedure suffers from the fact that it typically doesn't converge to ($\hat{\alpha}$,$\hat{\beta}$) that maximizes the likelihood.

3.4 Estimating Random Effects Covariance Matrix

If the covariance matrices, G and R_d, are unknown, but their estimates, say, G and R_d, are available. The estimates of β and $\alpha(.)$ thus can be obtained by substitution of G and R_d in (3.14) and (3.15). In this section, we consider the estimation of random effects covariance matrix G.

Recall that $\beta = (\beta_1^T, \beta_2^T)^T$, $\beta_d = (\beta_{d1}^T, \dots, \beta_{dn_c}^T)^T$, where $\{(\beta_{1i}^T, \beta_{2i}^T)^T, i = 1, \dots, n_c\}$

are i.i.d. $N(0, D)$. D is the $2q \times 2q$ covariance matrix of $(\beta_{1i}^T, \beta_{2i}^T)^T$. Suppose

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}.$$

Then

$$G = \begin{pmatrix} I_{n_c} \otimes D_{11} & I_{n_c} \otimes D_{12} \\ I_{n_c} \otimes D_{21} & I_{n_c} \otimes D_{22} \end{pmatrix}.$$

The unknown components of G can then be obtained using D. Restricted maximum likelihood (REML) is widely used to estimate variance components. However, the optimization problem associated with REML can be difficult for multilevel model with random effects. Empirical results suggest that the estimation procedures commonly implemented using REML may have difficulties in optimization when the dimension of the cluster effect is greater than 3, see Verbeke and Molenberghs (2000). We propose to use the method of moment estimation of D,

$$\hat{D} = \frac{1}{n_c} \sum_{i=1}^{n_c} \begin{pmatrix} \hat{\beta}_{1i} \\ \hat{\beta}_{2i} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{1i}^T & \hat{\beta}_{2i}^T \end{pmatrix}. \tag{3.16}$$

In the case of finite sample, there is no guarantee that $\hat{D}$ will be a positive-definite matrix. When $\hat{D}$ is not positive-definite, a shrinkage estimation of D,

$$\hat{D}_\lambda = \lambda \hat{D} + (1 - \lambda)\text{diag}(\hat{D}),$$

where $\lambda \in (0, 1)$, could be considered.

We can see that the estimate for G depends on the estimation of β. On the other hand, improving the efficiency of the estimation of (α, β) relies on the estimates for G and R. We further develop the estimation procedure for the parameters of error terms in the next section.

3.5 Estimating Subject Correlation Parameters

As mentioned in Section 3.4, the efficiency of the estimated coefficients relies on the estimates for D and R_d. In this section, we study the estimation of R_d.

Recall that $\epsilon_{dijk} = \epsilon_{dij}(t_k), k = 1, \ldots, n_m$, which is an Ornstein-Uhlenbeck process satisfing (2.11) as follows,

$$\epsilon_{dij}(t) = \epsilon_{dij}(0)e^{-\theta_d t} + \sigma_d \int_0^t e^{-\theta_d (t-s)}\, dW(s),$$

where $\epsilon_{dij}(0) \sim N(0, \frac{\sigma_d^2}{2\theta_d})$, $\theta_d > 0$ and $\sigma_d > 0$ are parameters, W_t is the Wiener process independent of $\epsilon_{dij}(0)$, and $t_k, k = 1, \ldots, n_m$, are in the range of interest, say, an interval $\mathcal{T} = [0, T]$. Assume that $0 \leqslant t_1 < \ldots, < t_m \leqslant T$. $\epsilon_{dij}(t_k), k = 1, \ldots, n_m$, have a Gaussian distribution with mean $E\epsilon_{dij}(t) = 0$ and variance $\mathrm{Var}(\epsilon_{dij}(t)) = \frac{\sigma_d^2}{2\theta_d}$, and correlation coefficients $\mathrm{Cov}(\epsilon_{dij}(t), \epsilon_{dij}(s)) = \frac{\sigma_d^2}{2\theta_d} e^{-\theta_d |t-s|}$. Let R_{ds} be the covariance matrix of $(\epsilon_{dij1}, \ldots, \epsilon_{dijn_m})^T$, then

$$R_{ds} = \frac{\sigma_d^2}{2\theta_d} \begin{pmatrix} 1 & e^{-\theta_d|t_1-t_2|} & e^{-\theta_d|t_1-t_3|} & \cdots \\ e^{-\theta_d|t_2-t_1|} & 1 & e^{-\theta_d|t_2-t_3|} & \cdots \\ e^{-\theta_d|t_3-t_1|} & e^{-\theta_d|t_3-t_2|} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and $R_d = I_{n_c n_s} \otimes R_{ds}$.

Note that for $s < t$, $\epsilon_{dij}(t)$ and $\epsilon_{dij}(s)$ are connected by

$$\epsilon_{dij}(t) = e^{-\theta_d(t-s)}\epsilon_{dij}(s) + \sigma_d \int_s^t e^{-\theta_d(t-u)}\, dW(u),$$

Let $\epsilon_{dij1}, \epsilon_{dij2}, \ldots, \epsilon_{dijn_m}$ be a sample of the Ornstein-Uhlenbeck process defined by (2.10), $r_{dij1}, r_{dij2}, \ldots, r_{dijn_m}$ are their observed values, the likelihood of $\epsilon_{dij1}, \ldots, \epsilon_{dijn_m}$, can be written as

$$L_{dij}(\theta_d, \sigma_d) = f_{\epsilon_{dij1},\ldots,\epsilon_{dijn_m}}(r_{dij1}, \ldots, r_{dijn_m}) \tag{3.17}$$

$$= f_{\epsilon_{dij1}}(r_{dij1}) \prod_{k=2}^{n_m} f_{\epsilon_{dijk}|\epsilon_{dij1}=r_{dij1},\ldots,\epsilon_{dij,k-1}=r_{dij,k-1}}(r_{dijk}|r_{dij1}, \ldots, r_{dij,k-1}) \tag{3.18}$$

$$= f_{\epsilon_{dij1}}(r_{dij1}) \prod_{k=2}^{n_m} f_{\epsilon_{dijk}|\epsilon_{dij,k-1}=r_{dij,k-1}}(r_{dijk}|r_{dij,k-1}) \tag{3.19}$$

$$= \prod_{k=1}^{n_m} \phi(r_{dijk}, \mu_k, \sigma_k), \tag{3.20}$$

where $\phi(r, \mu, \sigma)$ denotes the probability density function for normal distribution with mean μ and standard deviation σ, and

$$\mu_1 = 0, \quad \sigma_1^2 = \frac{\sigma_d^2}{2\theta_d},$$

$$\mu_k = e^{-\theta_d(t_k - t_{k-1})} r_{dij,k-1}, \quad k > 1,$$

$$\sigma_k^2 = (1 - e^{-2\theta_d(t_k - t_{k-1})}) \frac{\sigma_d^2}{2\theta_d}, \quad k > 1.$$

The third equality (3.19) is due to the Markovian property and the last equality (3.20) holds because of Gaussian property. After we get an estimate of β_d and α_d, we can calculate the residual and plug the residuals into the likelihood function. The maximum likelihood estimator for $\hat{\theta}_d$ and $\hat{\sigma}_d$ can be easily found by numerically maximizing the log-likelihood function (3.20).

3.6 Profile Likelihood Estimation Procedure

The estimation procedure includes four steps:

1. Find initial guesses for R_d, $d = 1, 2$ and G;
2. With initial guesses, estimate $\hat{\beta}$;
3. Use $\hat{\beta}$ to further estimate G and R_d;
4. Estimate $\hat{\beta}$ more efficiently by using the estimate for G and R_d;
5. Calculate $\hat{\alpha}_d$.

The estimation procedure utilizes

$$G = I_{2qn_c}, R_1 = R_2 = I_{n_c n_s n_m}$$

in the first step. Then using the G and R_d from step 1, we obtain an intial estimate for $(\hat{\beta}, \hat{\alpha})$ according to (3.14) and (3.15) in the step 2,

$$\hat{\beta} = \left[\begin{pmatrix} Z^T(I - S_1^T)(I - S_1)Z & \\ & Z^T(I - S_2^T)(I - S_2)Z \end{pmatrix} + I_{2qn_c} \right]^{-1} \begin{pmatrix} Z^T(I - S_1^T)(I - S_1)Y_1 \\ Z^T(I - S_2^T)(I - S_2)Y_2 \end{pmatrix}.$$

Move to the next step, we further estimate

$$\hat{D} = \frac{1}{n_c} \sum_{i=1}^{n_c} \begin{pmatrix} \hat{\beta}_{1i} \\ \hat{\beta}_{2i} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{1i}^T & \hat{\beta}_{2i}^T \end{pmatrix} =: \begin{pmatrix} \hat{D}_{11} & \hat{D}_{12} \\ \hat{D}_{21} & \hat{D}_{22} \end{pmatrix},$$

and

$$\hat{G} = \begin{pmatrix} I_{n_c} \otimes \hat{D}_{11} & I_{n_c} \otimes \hat{D}_{12} \\ I_{n_c} \otimes \hat{D}_{21} & I_{n_c} \otimes \hat{D}_{22} \end{pmatrix}.$$

The residuals are

$$\begin{pmatrix} \hat{E}_1 \\ \hat{E}_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} \hat{m}_1 \\ \hat{m}_2 \end{pmatrix} - \begin{pmatrix} Z & \\ & Z \end{pmatrix} \hat{\beta},$$

where

$$\begin{pmatrix} \hat{m}_1 \\ \hat{m}_2 \end{pmatrix} = \begin{pmatrix} S_1 & \\ & S_2 \end{pmatrix} \left[\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} Z & \\ & Z \end{pmatrix} \hat{\beta} \right].$$

Use the residuals to numerically find the MLE of θ_d and σ_d using (3.20), and construct $\hat{R}_d$.

In the final step, update $\hat{\beta}$ by

$$\hat{\beta} = V^{-1} \begin{pmatrix} Z^T(I - S_1)^T \hat{R}_1^{-1}(I - S_1)Y_1 \\ Z^T(I - S_2)^T \hat{R}_2^{-1}(I - S_2)Y_2 \end{pmatrix},$$

where

$$V = \begin{pmatrix} Z^T(I - S_1)^T \hat{R}_1^{-1}(I - S_1)Z & \\ & Z^T(I - S_2)^T \hat{R}_2^{-1}(I - S_2)Z \end{pmatrix} + \hat{G}^{-1}.$$

Estimate D by

$$\hat{D} = \frac{1}{n_c} \sum_{i=1}^{n_c} \begin{pmatrix} \hat{\beta}_{1i} \\ \hat{\beta}_{2i} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{1i}^T & \hat{\beta}_{2i}^T \end{pmatrix}.$$

Estimate α_d by

$$\hat{\alpha}_d(t) = (s_0 - s_1 s_2^{-1} s_1)^{-1} \times \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} (I_p - (t_{ijk} - t) s_1 s_2^{-1}) K_h(t_{ijk} - t) x_{ijk} (y_{dijk} - z_{dijk}^T \hat{\beta}_{di}).$$

The bandwidth h_d, $d = 1, 2$, are selected using cross validation.

3.7 Asymptotic Theory

Before we state the asymptotic properties of the proposed estimators, first we list the regularity conditions needed for the consistency and asymptotic normality of our estimators.

(1) $f(t)$ is continuously differentiable on $[0, T]$.

(2) $\inf_t f(t) > 0$.

(3) $\inf_t \det(\mathbf{E}x(t)x(t)^T) > 0$.

(4) $\alpha_d(t)$ is twice continuously differentiable.

(5) $\mathbf{E}\sup_t [x^l(t)]^2 < \infty, 1 \leqslant l \leqslant p$.

(6) $\mathbf{E}\sup_t [z^l(t)]^2 < \infty, 1 \leqslant l \leqslant q$.

(7) $\mathbf{E}\sup_t [\epsilon_d(t)]^2 < \infty$, $d = 1, 2$.

(8) The partial derivatives of $C_{x,z}$, $C_{x,\epsilon}$, and $C_{z,\epsilon}$ exist and are uniformly continuous on $[0, T]^2$ except on the diagonal line, where $C_{x,z}(u, v) = \mathbf{E}x(u)z(v)^T$.

(9) $\mathbf{E}x(t)x(t)^T$ and $\mathbf{E}x(t)z(t)^T$ have bounded derivatives, and $\psi_{z_{\perp x}, z_{\perp x}}(t)$ is invertible, where $\psi_{x,z}(t) = \lim\limits_{\substack{u>v \\ u,v \to t}} \frac{\mathbf{E}[x(u)-x(v)][z(u)-z(v)]^T}{u-v}$.

(10) The kernel function K is compactly supported, positive, continuously differentiable, and $\int uK(u)\,du = 0$.

(11) $\forall 0 \leqslant s, t \leqslant T$, $\mathbf{E}(\epsilon_d(s) | x(t), z(t)) = 0$, $d = 1, 2$.

Theorem 3.1. *Under regularity conditions (1)-(10), there exists a function* g *increasing to infinity, such that as* $n_c, n_s, n_m \to \infty$, $n_c^4 = o(n_m)$, $n_m = O(g(n_s))$, $h_d \to 0+$, $h_d {n_m}^{1/2} > \delta$ *for some* $\delta > 0$, $d = 1, 2$, *we have,*

$$\sqrt{n_c}\left\{\text{vec}(\hat{D}) - \text{vec}(D) + o_p(h_1^2) + o_p(h_2^2)\right\} \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where $\Sigma = \text{Var}(\text{vec}(\beta_i \beta_i^T))$, $\beta_i = (\beta_{1i}^T, \beta_{2i}^T)^T, i = 1, \dots, n_c$.

The pointwise asymptotic behaviour of $\hat{\alpha}_d(t)$ is as follows and its proof is provided in the appendix.

Theorem 3.2. *Under regularity conditions (1)-(11), for any* $t \in (0, T)$, *there exists a function* g *increasing to infinity, such that as* $n_s, n_m \to \infty$, $n_c^4 = o(n_m)$, $n_m = O(g(n_s))$, $h_d \to 0+$, $h_d {n_m}^{1/2} > \delta$ *for some* $\delta > 0$, $d = 1, 2$, *we have*

$$\sqrt{n_c n_s}\{\hat{\alpha}_d(t) - \alpha_d(t) - \textit{bias}\} \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where

$$\begin{aligned} \textit{bias} =& \frac{1}{2}\alpha_d''(t)h_d^2 + h_d^2[Ex(t)x^T(t)]^{-1}[Ex(t)z^T(t)][E\psi_{z_{\perp x}, z_{\perp x}}]^{-1}E\psi_{z_{\perp x}, u_d} \\ &+ o_p(h_1^2) + o_p(h_2^2), \\ \Sigma =& [Ex(t)x(t)^T]^{-1}E[x(t)x^T(t)\epsilon^2(t)][Ex(t)x(t)^T]^{-1}. \end{aligned}$$

3.8 Hypothesis Test

For varying coefficient models, it is very natural to ask whether some of the coefficient functions are actually constants. Specifically, denote by L an index set, the hypothesis we would like to test is

$$H_0 : \forall l \in L, \alpha_{dl} \text{ are constant } \longleftrightarrow H_1 : \exists l \in L, \alpha_{dl} \text{ is not constant.}$$

Following Cai, Fan and Yao (2000), we propose a test based on the comparison of the residual sum of squares(RSS). Let RRS_0 and RRS_1 be the residual sum of squares (RSS) under null and alternative hypothesis, respectively.

$$RSS = \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} (y_{ijk} - x_{ijk}^T \hat{\alpha}_d(t_{ijk}) - z_{ijk}^T \hat{\beta}_{di})^2$$

The test statistics is

$$T = (RRS_0 - RRS_1)/RRS_1. \tag{3.21}$$

If the null hypothesis is false, we expect the test statistic T to be large. To get an approximation to the distribution of T under the null hypothesis, we adopt a nonparametric bootstrap strategy. According to Field and Welsh (2007) and Ren et al. (2010), for hierarchical models without serial correlations, bootstrapping on the highest level without replacement and sampling other levels nested in the highest level with replacement works best among other nonparametric bootstrapping strategies. To keep the serial correlation structure, we borrow the idea of block bootstrap (Hall et cl. 1995) and adjust the bootstrapping strategy by sampling the error terms on the subject level. First, we fit the full model, and obtain the random effects $\hat{\beta}_{di}$ for cluster i and residual vector $\hat{\epsilon}_{dij}$ for subject j; next we fit the null model, and obtain $\hat{\alpha}_d$; finally, use $\hat{\beta}_{di}$, $\hat{\epsilon}_{dij}$ and $\hat{\alpha}_d$ to generate bootstrap samples, and calculate T for each sample. Thus we get an approximation to the distribution of T conditioned on the observed covariates under null hypothesis. In summary, the following bootstrap strategy is adopted:

1. Fit the full model to obtain $\hat{\beta}_i$ and the residual vector $\hat{\epsilon}_{dij}$.

2. Fit the null model to obtain an estimation of the varying coefficients $\hat{\alpha}_d$.

3. Sample β_i^* from $\{\hat{\beta}_i, i = 1, \dots, n_c\}$ without replacement. Sample ϵ_{dij}^* from $\{\hat{\epsilon}_{dij}, i = 1, \dots, n_c, j = 1, \dots, n_s\}$ with replacement. Then generate the bootstrap sample data by
$$y_{dijk}^* = x_{ijk}^T \hat{\alpha}_d + z_{ijk}^T \beta_{di}^* + \epsilon_{dijk}^*.$$

4. Calculate T^* based on the generated bootstrap sample $\{x_{ijk}, z_{ijk}, y_{dijk}^*\}$.

5. Repeat step 3-4 B times and get $T_1^*, \dots, T_B^*$;

6. The p-value for H_0 can be estimated as $\sum_{b=1}^{B} I\{T_b^* \geqslant T\}/B$.

Chapter 4

Comparison Studies of Models

To explore the finite sample performance of the estimation procedure, four major simulation studies are conducted. In Section 4.1, we investigate the influence of ignoring between response correlation on estimation. The main conclusion is that it is better to include between response correlation associated with random effects, but not to include that associated with error term into the model. In Section 4.2, we further investigate the benefit of modeling two responses together by comparing our model with univariate varying coefficient models and mixed effects models, both of which are special cases of our model. In Section 4.3, we investigate how the serial correlation in covariate impact the estimation of varying coefficients, and point out that incorrectly assuming temporally independence will result in underestimated bias and variance.

4.1 Between Response Correlation

In this simulation study, we investigate finite sampling property of estimation if we ignore between-response dependence of random effects or error terms in the estimation. We consider $n_c = 20$ clusters, $n_s = 10$ subjects in each cluster, and $n_m = 30$ measurements for each subject. The data generating process is

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_{11}(t_k)x_{ijk1} + \alpha_{12}(t_k)x_{ijk2} \\ \alpha_{21}(t_k)x_{ijk1} + \alpha_{22}(t_k)x_{ijk2} \end{pmatrix} + \begin{pmatrix} \beta_{1i1}z_{ijk1} + \beta_{1i1}z_{ijk2} \\ \beta_{2i1}z_{ijk} + \beta_{2i2}z_{ijk2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix}.$$

The nonparametric components are

$$\alpha_{11}(t) = \sin(2\pi \frac{t}{n_m}), \quad \alpha_{12}(t) = \cos(2\pi \frac{t}{n_m}),$$
$$\alpha_{21}(t) = \cos(2\pi \frac{t}{n_m}), \quad \alpha_{22}(t) = \sin(2\pi \frac{t}{n_m}).$$

The covariance matrix of normally distributed random effects are

$$D = 0.1 \begin{pmatrix} 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 \end{pmatrix}.$$

The error term satisfies

$$\mathrm{Var}(\epsilon_{dij}(t)) = 1, \quad \mathrm{Cov}(\epsilon_{1ij}(t), \epsilon_{2ij}(s)) = 0.4^{|t-s|},$$
$$\mathrm{Cov}(\epsilon_{1ij}(t), \epsilon_{2ij}(t)) = 0.6.$$

500 datasets are generated. Performances of estimation methods under different dependence assumptions are compared based on the 500 datasets. The estimation

methods under investigation include:

i) Ignore between-response correlation associated with random effects and error terms.

ii) Only ignore between-response correlation associated with error terms.

iii) Only ignore between-response correlation associated with error terms, shrinkage tuning parameter $\lambda = 0.6$.

iv) Take all correlations into account.

v) Take all correlations into account, shrinkage tuning parameter $\lambda = 0.6$.

The bandwidth of kernel smoothing is set to be $n_m/10$. To measure how good a predictor of β is, suppose $\hat{\beta}^b$ is the predictor for β^b based on the bth dataset, $b = 1, \ldots, 500$. Define $\mathrm{SSE}(\hat{\beta}^b)$ to be the sum of squared prediction errors of $\hat{\beta}^b$,

$$\mathrm{SSE}(\hat{\beta}^b) = \sum_{i=1}^{n_c} ||\hat{\beta}_i^b - \beta_i^b||^2$$

Using $\mathrm{SSE}(\hat{\beta}^b)$ as a measure of how accurate $\hat{\beta}^b$ predict β^b, the simulation result shows that: when between-response correlation associated with error terms are taken into account, the prediction accuracy will drop significantly, even worse than completely ignoring all between-response correlations; if between-response correlation associated with error terms is ignored, no-shrinkage is better than ignoring all between-response correlations, which in turn is better than shrinkage estimator. Figure 4.1 demonstrates the simulation result. In Figure 4.1, the sum of squared prediction error is compared with the one when responses are modelled

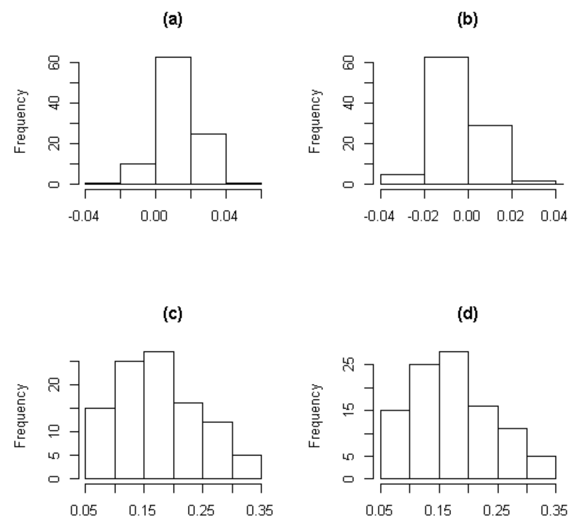

Figure 4.1: Histogram of the difference between squared prediction errors of two methods. In each panel, a particular method is compared with the method where responses are treated separately. Panel (a): ignore between-response correlation associated with error terms, and estimate D with shrinkage estimator. Panel(b): ignore between-response correlation associated with error terms. Panel (c): take all between-response correlation into account, and estimate D with shrinkage estimator. Panel(d): take all between-response correlation into account.

separately, and the difference is summarized in four histograms. However, the improvement of estimation accuracy is not significant for α. Figure 4.2 shows the estimations of $\alpha_{11}(t)$, $\alpha_{12}(t)$, $\alpha_{21}(t)$, $\alpha_{22}(t)$ in two cases: when two responses are considered simultaneously but the between-response correlation associated with error terms are ignored (left column); and when responses are modelled separately (right column). The t-test of squared estimation errors doesn't report significant difference. There are two possible reasons for the insignificant result: one is that α is estimated for each response separately in both cases, so there is no significant

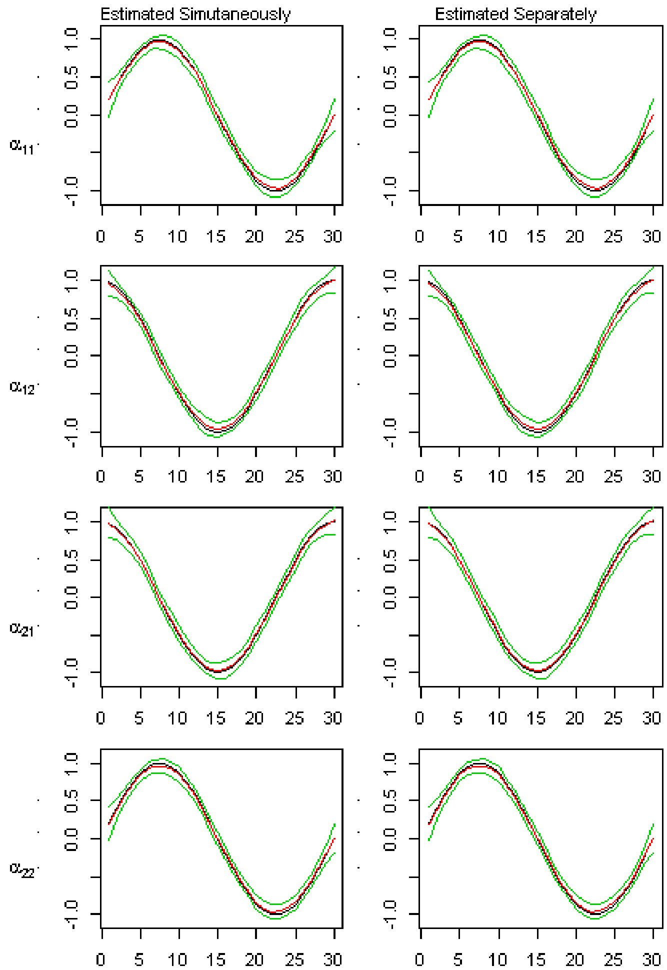

Figure 4.2: Black line: true value. Red line: mean of estimated value in 20 simulations. Green line: 99% point-wise confidence band.

improvement of α's accuracy; the other is that the improvement of β's estimation in the simulation is not large enough to make a statistically significant improvement of α's estimation.

4.2 Some Special Cases

We use simulation to demonstrate the performance of our model when the sample size is small. First, we compare the joint modeling with separate modeling when there are two responses. Second, we compare our model with the classic varying coefficient model using R package “np". Finally, we compare our model with the linear mixed effects model.

4.2.1 Univariate Model

In this simulation study, we demonstrate the benefit of modeling two responses simultaneously in terms of bias and variance of $\hat{D}$. Consider the model

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_1(t_k)x_{ijk} \\ \alpha_2(t_k)x_{ijk} \end{pmatrix} + \begin{pmatrix} \beta_{1i}z_{ijk} \\ \beta_{2i}z_{ijk} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix}.$$

In this model, x_{ijk} and z_{ijk} are univariate random variables, and $x_{ijk} = x_{ij}(t_k)$, $z_{ijk} = z_{ij}(t_k)$, where x_{ij}, z_{ij}, $i = 1, \ldots, n_c$, $j = 1, \ldots, n_s$, are independent Brownian motions, which satisfies $x_{ij}(0) = 0$, $z_{ij}(0) = 0$, $\mathrm{Var}(x_{ij}(t)) = t$, $\mathrm{Var}(z_{ij}(t)) = t/25$. The error term $\epsilon_{dijk} = \epsilon_{dij}(t_k)$, where ϵ_{dij}, $d = 1, 2$, $i = 1, \ldots, n_c$, $j = 1, \ldots, n_s$, are independent Ornstein-Uhlenbeck processes,

$$\epsilon_{dij}(t) = \epsilon_{dij}(0)e^{-\theta_d t} + \sigma_d \int_0^t e^{-\theta_d(t-s)}\, dW(s).$$

The parameter σ_d controls the variance of stochastic process, while the parameter θ_d controls the serial dependence of the process. σ_d and θ_d are chosen in the way such that $\mathrm{Var}(\epsilon_{dij}(t)) = 1$, $\mathrm{Cov}(\epsilon_{dij}(t), \epsilon_{dij}(s)) = 0.2^{|t-s|}$. Time points, $t_1, \ldots, t_{n_m}$

are randomly sampled from a uniform distribution on $[0,1]$. The varying coefficients are $\alpha_1(t) = \sin(t\pi)$, $\alpha_2(t) = \cos(t\pi)$. The random effects $(\beta_{1i}, \beta_{2i})^T$, $i = 1, \ldots, n_c$, are i.i.d. bivariate normal random vectors, with $\mathbf{E}\beta_{1i} = \mathbf{E}\beta_{2i} = 0$,

$$\mathrm{Var}((\beta_{1i}, \beta_{2i})^T) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\rho = 0.2$, 0.4, 0.6 or 0.8. We consider $n_c = 8$ or 16 clusters, $n_s = 20$ or 40 subjects in each cluster, and $n_m = 50$ measurements for each subjects. So there is a total of $4 \times 2 \times 2 = 16$ scenarios. For each scenario, 1000 datasets are generated. We estimate D based on each dataset through joint modeling and separate modeling, the latter of which ignores the correlation between the random effects β_{1i} and β_{2i} associated with the two responses. The bias and variance of $\hat{D}$ based on the 1000 datasets in each of the 16 scenarios are summarized in Tables 4.1 and 4.2 for joint modeling and separate modeling respectively.

We can draw several conclusions from Table 4.1. First, if the number of clusters n_c is doubled, the standard deviation of $\hat{D}$ is roughly decreased by a factor of $\sqrt{2}$, consistent with Theorem 3.1. Second, if the number of subjects n_s is doubled, the result is mixed, in the sense that standard deviations can increase as well as decrease.

Comparing Tables 4.1 and 4.2, we can see that in all of the 16 scenarios, the magnitude of bias is smaller for joint modeling than for separate modeling, especially for large ρ. The standard deviations of $\hat{D}_{11}$ and $\hat{D}_{22}$ are also smaller for joint modeling than for separate modeling, especially for large ρ. The standard deviation of $\hat{D}_{12}$ is smaller for separate modeling than for joint modeling, because separate modeling assumes D_{12} to be 0, and tends to produce a smaller estimate of D_{12} relative to joint

	ρ	$n_c = 8$					$n_c = 16$				
		$n_s = 20$		$n_s = 40$		ASE	$n_s = 20$		$n_s = 40$		ASE
		Bias(S.E.)	SD	Bias(S.E.)	SD		Bias(S.E.)	SD	Bias(S.E.)	SD	
$\hat{D}_{11}$	0.2	-0.06 (0.02)	0.55	-0.07 (0.02)	0.52	0.50	-0.05 (0.01)	0.38	-0.04 (0.01)	0.36	0.35
	0.4	-0.09 (0.02)	0.54	-0.04 (0.02)	0.56		-0.06 (0.01)	0.40	-0.04 (0.01)	0.38	
	0.6	-0.05 (0.02)	0.55	-0.05 (0.02)	0.54		-0.07 (0.01)	0.38	-0.01 (0.01)	0.39	
	0.8	-0.04 (0.02)	0.57	-0.04 (0.02)	0.53		-0.04 (0.01)	0.37	-0.07 (0.01)	0.24	
$\hat{D}_{22}$	0.2	-0.07 (0.02)	0.55	-0.04 (0.02)	0.52	0.50	-0.07 (0.01)	0.37	-0.04 (0.01)	0.37	0.35
	0.4	-0.08 (0.02)	0.52	-0.05 (0.02)	0.54		-0.10 (0.01)	0.36	-0.03 (0.01)	0.37	
	0.6	-0.04 (0.02)	0.55	-0.04 (0.02)	0.53		-0.09 (0.01)	0.36	-0.04 (0.01)	0.37	
	0.8	-0.03 (0.02)	0.55	-0.06 (0.02)	0.53		-0.07 (0.01)	0.36	-0.07 (0.01)	0.23	
$\hat{D}_{12}$	0.2	-0.01 (0.01)	0.39	-0.01 (0.01)	0.39	0.39	0.01 (0.01)	0.27	0.01 (0.01)	0.27	0.28
	0.4	-0.02 (0.01)	0.41	-0.01 (0.01)	0.40	0.44	-0.01 (0.01)	0.28	0.00 (0.01)	0.28	0.31
	0.6	0.01 (0.01)	0.46	-0.02 (0.01)	0.44	0.50	-0.01 (0.01)	0.30	0.01 (0.01)	0.31	0.35
	0.8	0.02 (0.02)	0.51	-0.02 (0.01)	0.47	0.55	0.00 (0.01)	0.33	0.01 (0.01)	0.21	0.39

Table 4.1: Joint modeling. Calculation of bias and standard deviations are based on 1000 simulations. Number of measurement for each individual is $n_m = 50$. ρ is correlation coefficient between random effects of the two responses. ASE is asymptotic standard error.

modeling. Table 4.3 combines bias and variance using mean squared error (MSE). $\hat{D}_{11}$ and $\hat{D}_{22}$ always have a smaller MSE under joint modeling, and the reduction in MSE increases as ρ increases.

Figure 4.3 demonstrates the average squared prediction error of $\hat{\beta}$ when $n_c = 30$, $n_s = 5$, and $n_m = 50$, based on 1000 datasets. The first row is separate modeling while the second row is joint modeling. The left column demonstrates the prediction error of $\hat{\beta}_{1i}$, and the right column demonstrates the prediction error of $\hat{\beta}_{2i}$. Each histogram is plotted in the same range. Comparing the first row with the second row, we can see that joint modeling reduces the prediction error of $\hat{\beta}$ by roughly 50%. Comparing the left column with the right one, there is no obvious difference between the prediction error for the random effects of the first response and that of the second response.

	ρ	$n_c = 8$				$n_c = 16$			
		$n_s = 20$		$n_s = 40$		$n_s = 20$		$n_s = 40$	
		Bias(S.E.)	SD	Bias(S.E.)	SD	Bias(S.E.)	SD	Bias(S.E.)	SD
$\hat{D}_{11}$	0.2	-0.06 (0.02)	0.56	-0.07 (0.02)	0.52	-0.05 (0.01)	0.38	-0.04 (0.01)	0.36
	0.4	-0.10 (0.02)	0.54	-0.04 (0.02)	0.56	-0.06 (0.01)	0.40	-0.04 (0.01)	0.38
	0.6	-0.06 (0.02)	0.55	-0.05 (0.02)	0.54	-0.07 (0.01)	0.38	-0.01 (0.01)	0.39
	0.8	-0.05 (0.02)	0.57	-0.04 (0.02)	0.53	-0.05 (0.01)	0.38	-0.07 (0.01)	0.24
$\hat{D}_{22}$	0.2	-0.07 (0.02)	0.55	-0.04 (0.02)	0.52	-0.07 (0.01)	0.37	-0.04 (0.01)	0.37
	0.4	-0.08 (0.02)	0.52	-0.05 (0.02)	0.54	-0.10 (0.01)	0.36	-0.03 (0.01)	0.37
	0.6	-0.04 (0.02)	0.56	-0.04 (0.02)	0.53	-0.09 (0.01)	0.36	-0.04 (0.01)	0.37
	0.8	-0.03 (0.02)	0.56	-0.06 (0.02)	0.53	-0.07 (0.01)	0.37	-0.07 (0.01)	0.24
$\hat{D}_{12}$	0.2	-0.03 (0.01)	0.36	-0.03 (0.01)	0.37	-0.02 (0.01)	0.24	-0.01 (0.01)	0.25
	0.4	-0.07 (0.01)	0.38	-0.03 (0.01)	0.38	-0.05 (0.01)	0.26	-0.02 (0.01)	0.27
	0.6	-0.06 (0.01)	0.43	-0.06 (0.01)	0.43	-0.08 (0.01)	0.29	-0.03 (0.01)	0.30
	0.8	-0.08 (0.02)	0.49	-0.07 (0.01)	0.46	-0.09 (0.01)	0.32	-0.11 (0.01)	0.20

Table 4.2: Separate modeling. Calculation of bias and standard deviations are based on 1000 simulations. Number of measurement for each individual is $n_m = 50$. ρ is correlation coefficient between random effects of the two responses.

4.2.2 Nonparametric Model

In this simulation study, we compare our model with existing varying coefficient models, using R package "np". The same kernel function, Epanechnikov kernel, is used in both methods. Since a ready-to-use R package for a model that incorporates both varying coefficients and random effects is not available, to make the comparison fair, we first consider a data generating process with no random effects,

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_{11}(t_k) + \alpha_{12}(t_k)x_{ijk} \\ \alpha_{21}(t_k) + \alpha_{22}(t_k)x_{ijk} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix}$$

where

- $x_{ijk} = x_{ij}(t_k)$, $\epsilon_{dijk} = \epsilon_{dij}(t_k)$;

MSE	ρ	$n_c = 8$				$n_c = 16$			
		$n_s = 20$		$n_s = 40$		$n_s = 20$		$n_s = 40$	
		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
$\hat{D}_{11}$	0.2	0.3120	0.3108	0.2790	0.2789	0.1506	0.1502	0.1288	0.1287
	0.4	0.3047	0.3034	0.3182	0.3173	0.1662	0.1653	0.1453	0.1451
	0.6	0.3112	0.3086	0.2937	0.2927	0.1510	0.1493	0.1511	0.1508
	0.8	0.3323	0.3245	0.2815	0.2792	0.1428	0.1394	0.0623	0.0598
$\hat{D}_{22}$	0.2	0.3055	0.3049	0.2729	0.2726	0.1455	0.1451	0.1383	0.1383
	0.4	0.2797	0.2783	0.2972	0.2966	0.1431	0.1423	0.1397	0.1395
	0.6	0.3105	0.3070	0.2806	0.2794	0.1390	0.1376	0.1382	0.1379
	0.8	0.3104	0.3045	0.2861	0.2833	0.1418	0.1380	0.0624	0.0600
$\hat{D}_{12}$	0.2	0.1278	0.1529	0.1372	0.1508	0.0601	0.0730	0.0643	0.0715
	0.4	0.1486	0.1696	0.1461	0.1577	0.0691	0.0787	0.0729	0.0789
	0.6	0.1917	0.2088	0.1859	0.1939	0.0891	0.0930	0.0914	0.0959
	0.8	0.2476	0.2598	0.2193	0.2231	0.1122	0.1111	0.0523	0.0439

Table 4.3: Mean squared error comparison. Column (1) is separate modeling. Coulumn (2) is joint modeling.

- $x_{ij}(t)$ is a standard Brownian motion;
- $\epsilon_{dij}(t)$ is an OU process, with $\text{Var}(\epsilon_{dij}(t)) = 1, \text{Cov}(\epsilon_{dij}(u), \epsilon_{dij}(v)) = 0.2^{|u-v|}$;
- x_{ij}, ϵ_{dij}, $d = 1, 2$, $i = 1, \dots, n_c$, $j = 1, \dots, n_s$, $k = 1, \dots, n_m$, are independent;
- t_k, $k = 1, \dots, n_m$, are independent uniform random variables;
- $\alpha_{11}(t) = \alpha_{22}(t) = \sin(t\pi)$, $\alpha_{12}(t) = \alpha_{21}(t) = \cos(t\pi)$.

We generate 500 datasets, each of which consists of $n_c = 5$ clusters, $n_s = 10$ subjects, and $n_m = 50$ time points. Bias and standard error associated with three bandwidths, namely $h = 0.10, 0.15$ and 0.20 are investigated. For the "np" package, we also include the optimal bandwidth automatically selected by the package. The

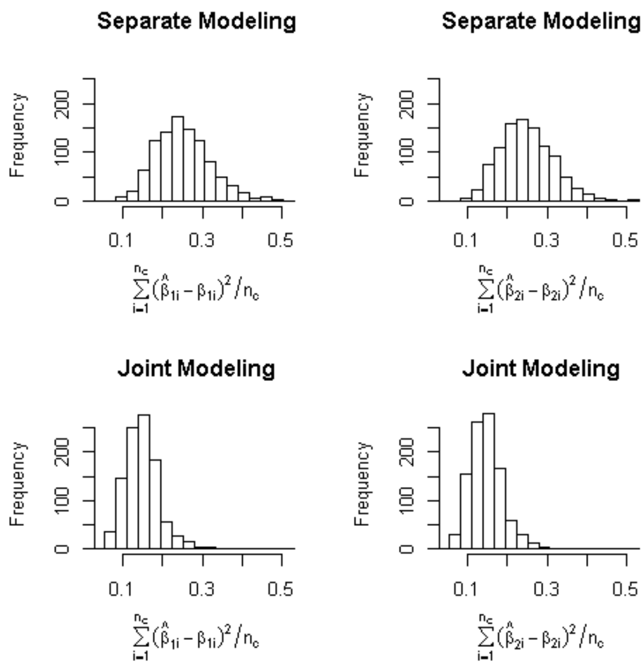

Figure 4.3: Histogram of prediction error of β. First row is separate modeling while second row is joint modeling. Left column demonstrates the prediction error of $\hat{\beta}_{1i}$, and right column demonstrates the prediction error of $\hat{\beta}_{2i}$. Each histogram is plotted in the same range.

result is demonstrated in Figure 4.4 for $\hat{\alpha}_{11}$ and in Figure 4.5 for $\hat{\alpha}_{12}$. We do not show the result for $\hat{\alpha}_{21}$ and $\hat{\alpha}_{22}$, since it resembles the result for $\hat{\alpha}_{12}$ and $\hat{\alpha}_{11}$, respectively.

First, we compare the bias and standard errors of $\hat{\alpha}_{11}$ estimated by our methods and the R package "np" in Figure 4.4. Keep in mind that $\alpha_{11}(t) = \sin(t\pi)$. The bias of our method is smaller than that of the package "np" by roughly one order of magnitude, whether we compare the biases associated with the same bandwidth or not. Since $\alpha_{11}(t)$ is basically linear near the endpoints 0 and 1, our method produces a significantly smaller bias toward the endpoints relative to the middle part of the interval $[0, 1]$, just as the theoretical result predicts. On the contrary, the "np" package produces larger bias towards the endpoints. Even for the optimal bandwidth, the bias near the endpoints is still pretty large. As the bandwidth increases, the biases of both methods increase, while the standard errors of both methods decrease, except in the range $[0, 0.2]$ for package "np". The standard errors of the two methods are roughly of the same order of magnitude, however, relative to package "np", our method has a smaller standard errors in most part of the interval $[0, 1]$. The theoretical asymptotic standard errors for our method, represented by the solid black line in the right bottom graph in Figure. 4.4, is $(n_c n_s)^{-1/2}$, and dominates the empirical standard errors curves based on the 500 datasets for h =0.10, 0.15 and 0.20. The spikes of the red dotted line in the right bottom graph indicates that $h = 0.1$ is too small.

Next, we compare the bias and standard errors of $\hat{\alpha}_{12}$ in Figure 4.5. We compare standard errors first, since the standard errors have a very interesting pattern. Recall that $\alpha_{12}(t) = \cos(t\pi)$ and $x_{ij}(t)$ is a standard Brownian motion, which implies that $\mathbf{E}(x_{ij}^2(t)) = t$. In particular, $x_{ij}(t) \approx 0$ for small t, which means that the effect of α_{12} is hard to detect near the left endpoint 0, and as a consequence, $\hat{\alpha}_{12}$ should tend to have a larger variance near 0 relative to in the other part on $[0, 1]$. Theoretically, the asymptotic standard deviation of $\hat{\alpha}_{12}$ for our method is $(t n_c n_s)^{-1/2}$, which fits the empirical standard errors curves pretty well in the right bottom graph of Figure 4.5. Again, it dominates the empirical standard errors curves. The bias of our method is one magnitude smaller than that of the "np" package. It increases in the middle part of $[0, 1]$, mainly because $\alpha_{12}(t) = \cos(t\pi)$ is convex near 0 and concave near 1.

It is unclear why the bias of $\hat{\alpha}_{12}(t)$ suddenly moves towards the opposite direction when t approaches the endpoints for both of the methods.

Figure 4.6 combines the results in Figure 4.4 and 4.5 by depicting the root mean squared error(RMSE), which is defined as the root of bias^2+ SE^2. A smaller RMSE implies a better estimation in the sense of quadratic loss. In this simulation study, it is clear from Figure 4.4 and 4.5 that the standard errors dominates bias, so the RMSE curves basically resemble the shape of standard errors curves. For $\hat{\alpha}_{11}$, the RMSE of our method is smaller than that of package "np" for the most part of $[0, 1]$.

However, it is not easy to tell from Figure 4.6 which method is better for $\hat{\alpha}_{12}$. Thus a global measure of precision is required to facilitate the comparison. We use the mean integrated squared error(MISE) to accomplish the task. In our case, the MISE is defined as $\mathbf{E}\int_0^1(\hat{\alpha}_d(t) - \alpha_d(t))^2\,dt$, elementwisely. We use the 500 datasets to estimate the MISE. The result is summarized in Table 4.4. For $\hat{\alpha}_{11}$, the MISE of our method is around 0.0179, independent of the choice of the bandwidth, while the best MISE of package "np" is 0.0198. For $\hat{\alpha}_{12}$, package "np" has a lower MISE. However, our model has a smaller MISE on the interval $[0.4, 1]$ compared with the interval $[0.4, 1]$ by package 'np'.

MISE	$\hat{\alpha}_{11}$		$\hat{\alpha}_{12}$	
	Proposed method	Package "np"	Proposed method	Package "np"
$h = 0.10$	0.0179 (0.0008)	0.0252 (0.0008)	0.2480 (0.1476)	0.0518 (0.0021)
$h = 0.15$	0.0173 (0.0008)	0.0394 (0.0009)	0.0755 (0.0030)	0.0675 (0.0025)
$h = 0.20$	0.0173 (0.0008)	0.0596 (0.0010)	0.0673 (0.0027)	0.1127 (0.0034)
optimal h		0.0198 (0.0008)		0.0588 (0.0023)

Table 4.4: Mean integrated squared error. No random effects are included in the data generating process.

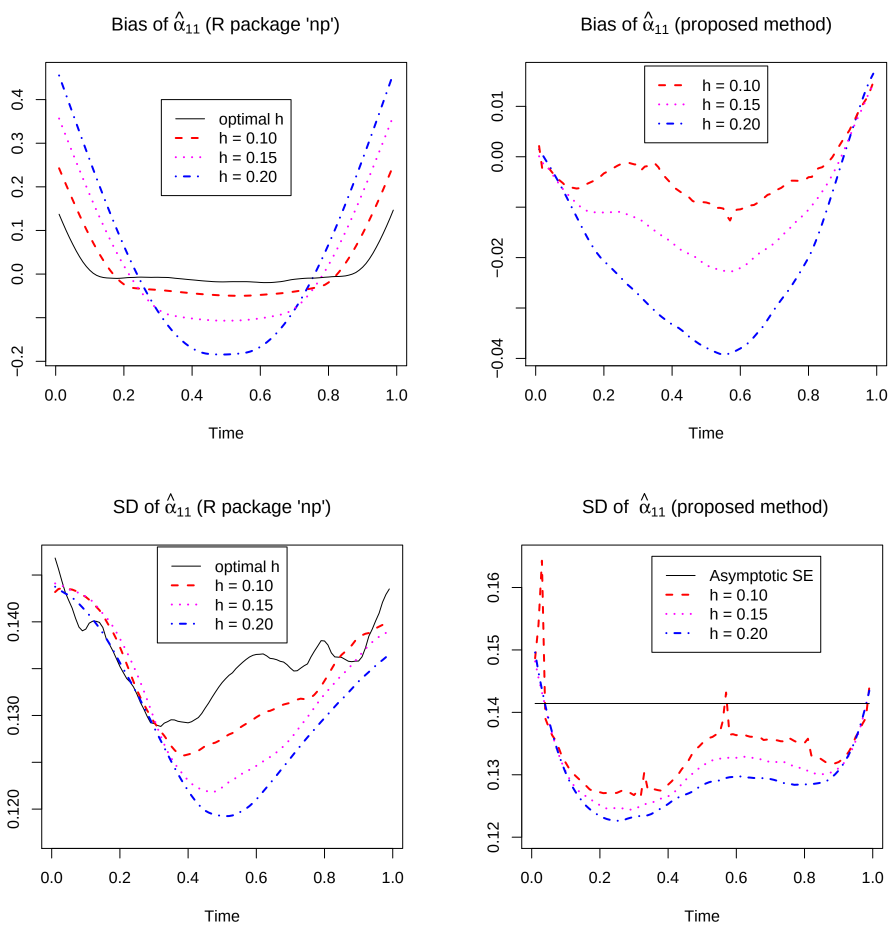

Figure 4.4: Comparison of pointwise bias and standard error of $\hat{\alpha}_{11}$ at different time points between 0 and 1. Left and right graphs show the result of "np" package and the proposed method respectively. Calculation is based on 500 datasets. Three bandwidths, namely 0.10, 0.15 and 0.20 are considered for both methods. For the "np" package, the bias and standard error associated with an additional bandwidth automatically selected by the package is also included (see the solid black line on the left panel). In the right bottom graph, the solid black line represents the theoretical asymptotic standard error.

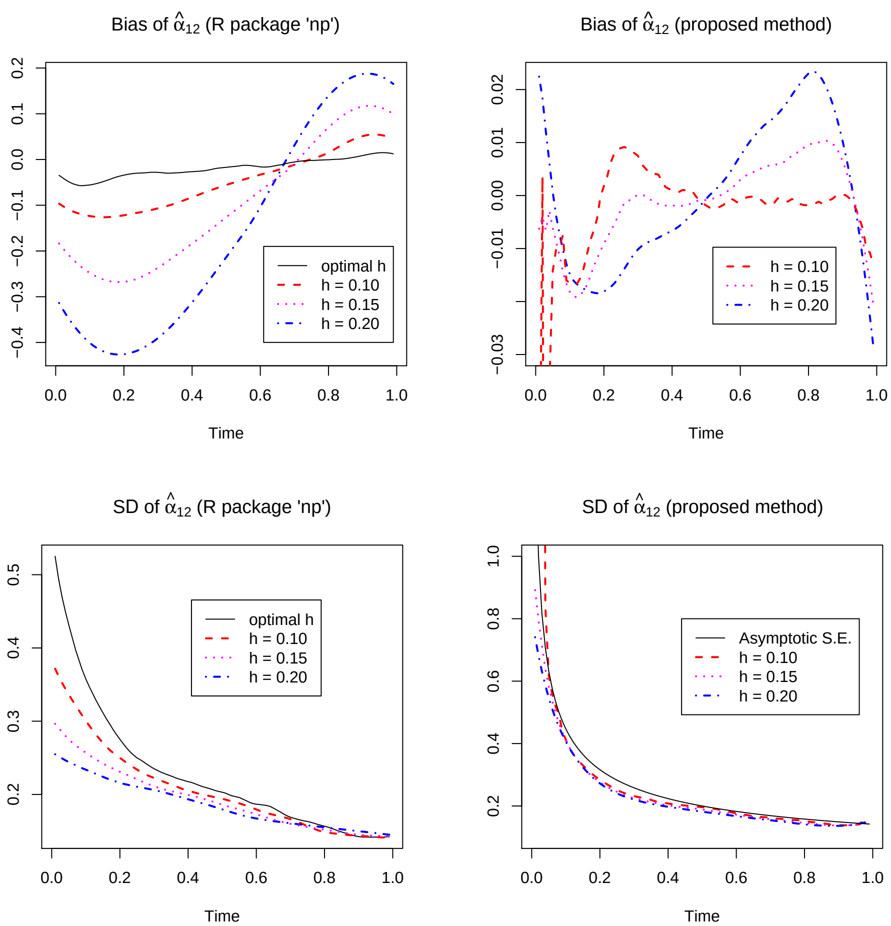

Figure 4.5: Comparison of pointwise bias and standard error of $\hat{\alpha}_{12}$ at different time points between 0 and 1. Left and right graphs show the result of "np" package and the proposed method respectively. Calculation is based on 500 datasets. Three bandwidths, namely 0.10, 0.15 and 0.20 are considered for both methods. For the "np" package, the bias and standard error associated with an additional bandwidth automatically selected by the package is also included (see the solid black line on the left panel). In the right bottom graph, the solid black line represents the theoretical asymptotic standard error.

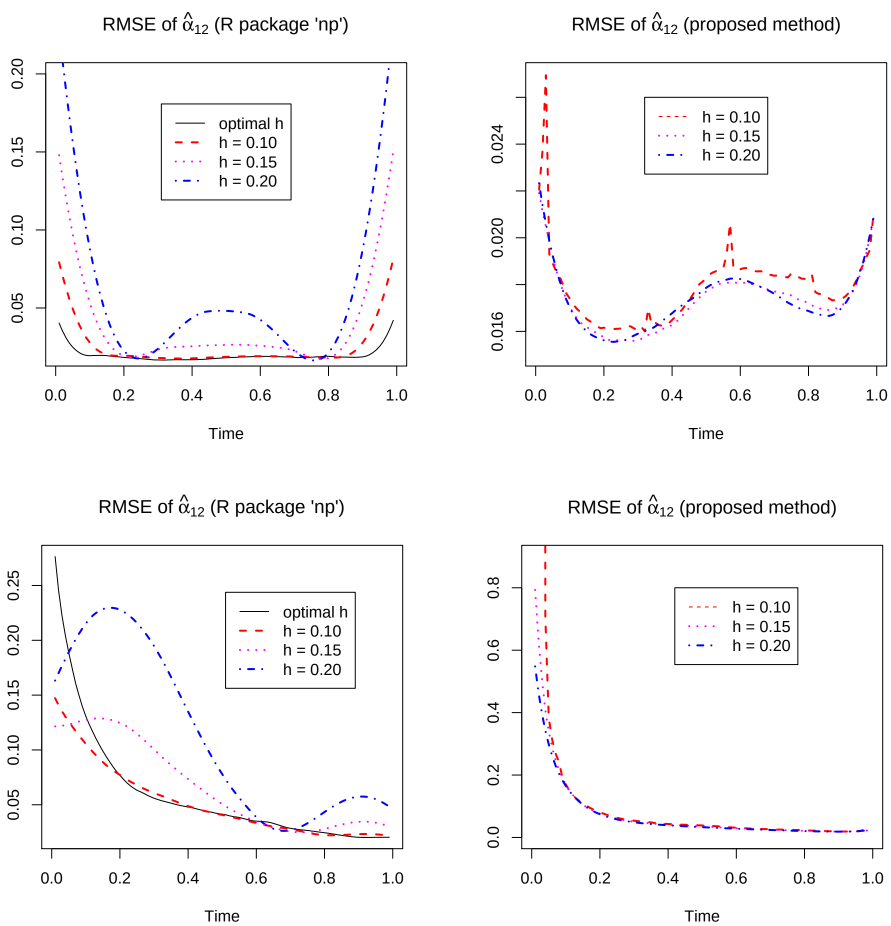

Figure 4.6: Comparison of pointwise root-mean-squared error of $\hat{\alpha}_{11}$ and $\hat{\alpha}_{12}$ at different time points between 0 and 1. Left and right graphs show the result of "np" package and the proposed method respectively. Calculation is based on 500 datasets. Three bandwidths, namely 0.10, 0.15 and 0.20 are considered for both methods. For the "np" package, the bias and standard error associated with an additional bandwidth automatically selected by the package is also included (see the solid black line on the left panel).

Next, we include the random effects and compare our model with the classic varying coefficient model. Assume the data is generated by

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_{11}(t_k) + \alpha_{12}(t_k)x_{ijk} \\ \alpha_{21}(t_k) + \alpha_{22}(t_k)x_{ijk} \end{pmatrix} + \begin{pmatrix} \beta_{1i}z_{ijk} \\ \beta_{2i}z_{ijk} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix},$$

where

$$\mathrm{Var}((\beta_{1i}, \beta_{2i})^T) = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

The MISE is summarized in Table 4.5. For $\hat{\alpha}_{11}$, the MISE of our method is around 0.0179, independent of the choice of the bandwidth, while the best MISE of package "np" is 0.0319, twice as large as our MISE. For $\hat{\alpha}_{12}$, except for $h = 0.1$, the largest MISE of our method is smaller than the smallest MISE of package "np".

MISE	$\hat{\alpha}_{11}$		$\hat{\alpha}_{12}$	
	Proposed method	Package "np"	Proposed method	Package "np"
$h = 0.10$	0.0179 (0.0008)	0.0365 (0.0014)	0.2487 (0.1481)	0.0698 (0.0032)
$h = 0.15$	0.0173 (0.0008)	0.0501 (0.0015)	0.0755 (0.0030)	0.0843 (0.0037)
$h = 0.20$	0.0173 (0.0008)	0.0699 (0.0016)	0.0673 (0.0027)	0.1286 (0.0046)
optimal h		0.0319 (0.0014)		0.0767 (0.0032)

Table 4.5: Mean integrated squared error. Random effects are included in the data generating process.

4.2.3 Linear Mixed Effects Model

In this simulation study, we compare our model with mixed effects models. Assume that the data is generated by

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_1(t_k)x_{ijk} \\ \alpha_2(t_k)x_{ijk} \end{pmatrix} + \begin{pmatrix} \beta_{1i}z_{ijk} \\ \beta_{2i}z_{ijk} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix},$$

where $\alpha_1(t) = \alpha_2(t) = (1+t)^2$, and

$$\text{Var}((\beta_{1i}, \beta_{2i})^T) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We consider two cases, $\rho = 0.8$ or 0. Set bandwidth equal to 0.5. Suppose the mixed effects model correctly specifies the linear structure,

$$y_{1ijk} = a_1 x_{ijk} + a_2 t_k x_{ijk} + a_3 t_k^2 x_{ijk} + b z_{ijk} + \epsilon_{1ijk}.$$

First we look at the fixed effects part. Let the estimation error of the mixed effects model be

$$\text{err}_k = \hat{a}_1 + \hat{a}_2 t_k + \hat{a}_3 t_k^2 - \alpha_1(t_k),$$

we calculate the overall bias

$$\frac{1}{n_m} \sum_{k-1}^{n_m} \text{err}_k,$$

and the standard deviation

$$\sqrt{\frac{1}{n_m} \sum_{k-1}^{n_m} (\text{err}_k - \overline{\text{err}})^2}.$$

The mean bias and standard deviations of the 1000 datasets are summarized in Table 4.6. The first row corresponds to dependent random effects, while the second row corresponds to independent random effects. Table 4.6 shows that the dependence structure does not influence the estimation of α very much in this simulation setting. Our method has a larger bias, but smaller standard deviations. The MSE is 0.0182 for our method and 0.0197 for mixed effects model. Now let's look at the random

	Proposed Model			Mixed effects Model		
h	bias	sd	MSE	bias	sd	MSE
0.5	0.0201	0.1334	0.0182	0.0113	0.1400	0.0197
0.5	0.0201	0.1331	0.0181	0.0113	0.1400	0.0197

Table 4.6: First row: no dependence between β_{1i} and β_{2i}. Second row: β_{1i} and β_{2i} are dependent.

effects. Figure 4.7 is the histogram plot for the mean squared prediction error of $\hat{\beta}_{1i}$. Because the prediction error of our method is about one tenth of the mixed effects model, the left column is not plotted in the same range as the right column. It is clear that our method has a better prediction of the random effects.

4.3 Influence of Covariates on Varying Coefficients Estimation

In this section, we focus on the influence of serial correlation of covariates on the proposed estimation method. Every setting includes 5 clusters, 10 subjects per cluster, and 50 measurements per subject. 1000 datasets are generated for each setting. The data generating process is

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \begin{pmatrix} \alpha_{10}(t_k) + x_{ijk1}\alpha_{11}(t_k) + x_{ijk2}\alpha_{12}(t_k) \\ \alpha_{20}(t_k) + x_{ijk1}\alpha_{21}(t_k) + x_{ijk2}\alpha_{22}(t_k) \end{pmatrix} + \begin{pmatrix} z_{ijk0}\beta_{1i0} + z_{ijk1}\beta_{1i1} + z_{ijk2}\beta_{1i2} \\ z_{ijk0}\beta_{2i0} + z_{ijk1}\beta_{2i1} + z_{ijk2}\beta_{2i2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ijk} \\ \epsilon_{2ijk} \end{pmatrix}.$$

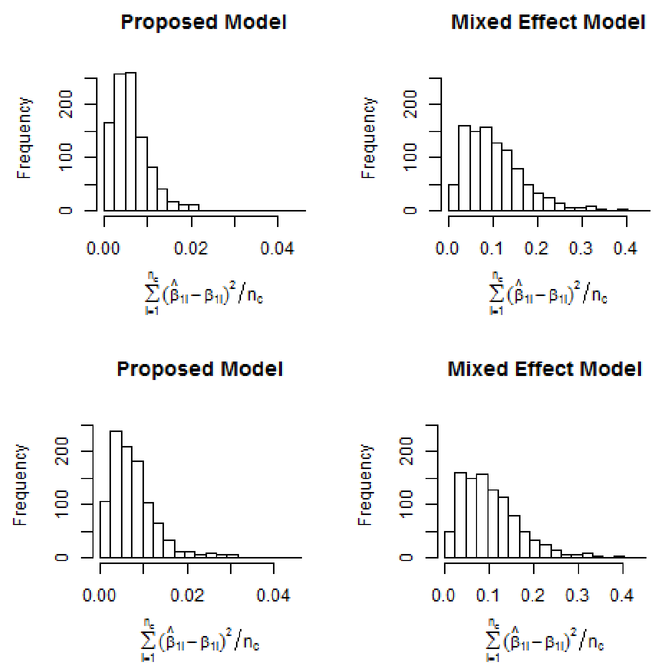

Figure 4.7: Histogram of prediction error of β. First row: no dependence between β_{1i} and β_{2i}. Second row: β_{1i} and β_{2i} are dependent.

In the model, time points are i.i.d. uniform distribution on $[0, 1]$. The nonparametric part is $\alpha_{10}(t) = \alpha_{20}(t) = 1$, $\alpha_{11}(t) = \alpha_{22}(t) = \sin(\pi t)$, $\alpha_{12}(t) = \alpha_{21}(t) = \cos(\pi t)$. For a given subject, z_{ijk0} is baseline measurement, and does not depend on time. z_{ijk0} is intended to mimic features that would not change with time, such as gender. There are two concerns about the simulation study. One concern is that the same covariates could appear in both varying coefficients component and random effects part. So we simulate $x_{ijk1} = z_{ijk1}$ or x_{ijk1} is independent of z_{ijk1}. The other concern is about the between-response correlation, which is controlled by the covariance matrix D of the random effects, and serial correlation in error terms. So we set

$$D = I_3 \otimes \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \mathrm{Cov}(\epsilon_{ij}(t), \epsilon_{ij}(s)) = \rho^{|s-t|},$$

where $\rho = 0.8$ or 0. Our primary interest is about the influence of serial correlation in covariates on the estimation. So we simulate x_{ijk1} and z_{ijk1} to be white noise or OU processes. x_{ijk2} and z_{ijk2} are white noises in all settings. A concise code for the settings is listed in the following:

hct: $\rho = 0.8$, $x_1 = z_1$ are OU process;

hcn: $\rho = 0.8$, $x_1 = z_1$ are white noise;

hit: $\rho = 0.8$, $x_1 \perp z_1$ are OU process;

hin: $\rho = 0.8$, $x_1 \perp z_1$ are white noise;

lct: $\rho = 0$, $x_1 = z_1$ are OU process;

lcn: $\rho = 0$, $x_1 = z_1$ are white noise;

lit: $\rho = 0$, $x_1 \perp z_1$ are OU process;

lin: $\rho = 0$, $x_1 \perp z_1$ are white noise.

The simulation results are summarized in Table 4.7 and 4.8. First, the table shows that in all settings, if the covariates have a serial correlation, $\hat{\alpha}_{11}$ will have a larger MISE. This result means that incorrectly assuming a covariate to be temporally independent will lead to an underestimate of the bias or variance of the estimator. Second, we can see that when ρ decreases, i.e. the correlation between responses and the serial correlation in error terms decrease, the MISE of $\hat{\alpha}_{11}$ decreases if x_1 is serially correlated, and increases if x_1 is white noise. Finally, we see that if x_1 appears in both varying coefficients part and random effects part, the MISE of $\hat{\alpha}_{11}$ will increase dramatically. This suggests that it is a good practice not to include the same variable in both varying coefficients part and random effects part of the model. Similar conclusion can be drawn for $\hat{\alpha}_{21}$.

h_1	ρ		$x_1 \perp z_1$		$x_1 = z_1$	
			Serial correlated	White noise	Serial correlated	White noise
0.15	0.8	$\hat{\alpha}_{10}$	0.0500 (0.0027)	0.0426 (0.0021)	0.0470 (0.0022)	0.0395 (0.0021)
		$\hat{\alpha}_{11}$	0.0285 (0.0012)	0.0029 (0.0001)	0.2272 (0.0103)	0.2083 (0.0096)
		$\hat{\alpha}_{12}$	0.0016 (0.0002)	0.0030 (0.0002)	0.0017 (0.0002)	0.0031 (0.0001)
	0.0	$\hat{\alpha}_{10}$	0.0036 (0.0005)	0.0034 (0.0005)	0.0035 (0.0005)	0.0034 (0.0004)
		$\hat{\alpha}_{11}$	0.0037 (0.0002)	0.0032 (0.0001)	0.2106 (0.0097)	0.2092 (0.0096)
		$\hat{\alpha}_{12}$	0.0032 (0.0001)	0.0031 (0.0001)	0.0032 (0.0001)	0.0032 (0.0001)
0.20	0.8	$\hat{\alpha}_{10}$	0.0595 (0.0035)	0.0573 (0.0034)	0.0549 (0.0027)	0.0498 (0.0032)
		$\hat{\alpha}_{11}$	0.0280 (0.0012)	0.0022 (0.0001)	0.2239 (0.0102)	0.2075 (0.0096)
		$\hat{\alpha}_{12}$	0.0007 (0.0000)	0.0023 (0.0001)	0.0008 (0.0000)	0.0022 (0.0001)
	0.0	$\hat{\alpha}_{10}$	0.0023 (0.0001)	0.0021 (0.0001)	0.0022 (0.0001)	0.0021 (0.0001)
		$\hat{\alpha}_{11}$	0.0032 (0.0001)	0.0029 (0.0001)	0.2096 (0.0097)	0.2085 (0.0096)
		$\hat{\alpha}_{12}$	0.0026 (0.0001)	0.0026 (0.0001)	0.0027 (0.0001)	0.0026 (0.0001)

Table 4.7: MISE of $\hat{\alpha}_1$

h_1	ρ		$x_1 \perp z_1$		$x_1 = z_1$	
			Serial correlated	White noise	Serial correlated	White noise
0.15	0.8	$\hat{\alpha}_{20}$	0.0459 (0.0022)	0.0379 (0.0018)	0.0529 (0.0031)	0.0448 (0.0024)
		$\hat{\alpha}_{21}$	0.0301 (0.0015)	0.0030 (0.0001)	0.2303 (0.0109)	0.1973 (0.0097)
		$\hat{\alpha}_{22}$	0.0016 (0.0001)	0.0030 (0.0001)	0.0018 (0.0002)	0.0030 (0.0001)
	0.0	$\hat{\alpha}_{20}$	0.0033 (0.0003)	0.0032 (0.0003)	0.0033 (0.0003)	0.0031 (0.0003)
		$\hat{\alpha}_{21}$	0.0033 (0.0001)	0.0034 (0.0002)	0.2040 (0.0091)	0.2053 (0.0091)
		$\hat{\alpha}_{22}$	0.0032 (0.0001)	0.0032 (0.0001)	0.0032 (0.0001)	0.0032 (0.0001)
0.20	0.8	$\hat{\alpha}_{20}$	0.0545 (0.0028)	0.0489 (0.0028)	0.0709 (0.0071)	0.0573 (0.0034)
		$\hat{\alpha}_{21}$	0.0302 (0.0014)	0.0022 (0.0001)	0.2300 (0.0109)	0.1968 (0.0097)
		$\hat{\alpha}_{22}$	0.0009 (0.0001)	0.0023 (0.0001)	0.0010 (0.0001)	0.0023 (0.0001)
	0.0	$\hat{\alpha}_{20}$	0.0022 (0.0001)	0.0021 (0.0001)	0.0022 (0.0001)	0.0021 (0.0001)
		$\hat{\alpha}_{21}$	0.0028 (0.0001)	0.0027 (0.0001)	0.2034 (0.0091)	0.2043 (0.0091)
		$\hat{\alpha}_{22}$	0.0030 (0.0001)	0.0030 (0.0001)	0.0030 (0.0001)	0.0030 (0.0001)

Table 4.8: MISE of $\hat{\alpha}_2$

Chapter 5

Real Data Analysis

The autoregressive integrated moving average (ARIMA) model is perhaps the most popular model to predict stock price. The drawback is that it can not easily pick up the nonlinear part of a stock price. Pai and Lin (2005) incorporated a support vector machine in an ARIMA model to capture the nonlinear pattern. They trained the model on the closing price of 10 stocks from Oct. 21, 2002 to Dec. 31, 2002, using January price as a validation dataset and February price as a test dataset. Hassan and Nath (2005) build Hidden Markov Model using opening price, closing price, highest price and lowest price in a duration of one and half a year to predict closing price next day for airline stocks. Hamao et al. (1990) analyzed correlations of three stock indexes in New York, London, and Tokyo by applying an autoregressive conditionally heteroskedasticity (ARCH) model on opening and closing price of these indexes over three year periods. Tsang et al. (2007) used artificial neural networks (ANN) to report a buy or sell signal. In addition to price variables, they also included trading volume and 5-day momentum, the price difference between now and 5 days ago, in their model.

5.1 Data

The daily stock data was downloaded from WRDS website (`https://wrds-web.wharton.upenn.edu/wrds`). The most recent NAICS Index File (version 2012) was downloaded from NAICS (`http://www.census.gov/eos/www/naics/2012NAICS/2012_NAICS_Index_File.xls`).

The stock data consists of daily information from 01/03/2006 - 12/31/2015. There are nine market holidays, including New Year's Day, so we do not have data on January 1st. There were 2943 stocks on the market on January 3nd 2006, and 2224 stocks on December 31, 2016. Delisted stocks are not considered. PERMNO (permanent number) is used to identify stocks, since the widely used Ticker symbols and CUSIP codes are subjected to change.

CRSP stock dataset covers 4 exchanges, including NYSE, AMEX, and the Nasdaq Stock Market. The types of shares traded in these exchanges include ordinary common shares, certificates, American depository receipts, etc. Only ordinary common shares (Share Code = 11) traded in Nasdaq Stock Market (Exchange Code = 3) are considered.

Table 5.1 summarizes relevant variables available from the data base.

Stocks can be grouped according to their business types. The grouping standard used by Federal statistical agencies is the North American Industry Classification System (NAICS). NAICS code is a 6-character code used to group companies with similar products or services. It was adopted in 1997 and implemented in 1999,

Variable	Description
EXCHCD	Exchange Code indicates the exchange on which a security is listed. Normal exchange codes are respectively 1,2, and 3 for NYSE, AMEX and the Nasdaq Stock MarketSM
SHRCD	Share Code. A two-digit code describing the type of shares traded. The first digit describes the type of security traded. 1 means "Ordinary Common Shares".
SPRTRN	Return on the Standard & Poor's Composite Index
NAICS	North American Industry Classification System Code
PRC	Prc is the closing price or the negative bid/ask average for a trading day. If the closing price is not available on any given trading day, the number in the price field has a negative sign to indicate that it is a bid/ask average.
VOL	Share Volume. VOL is the total number of shares of a stock sold on day I. It is expressed in units of one share, for daily data.
ASKHI	Ask or High Price is the highest trading price during the day, or the closing ask price on days when the closing price is not available. The field is set to zero if no Ask or High Price is available.
BIDLO	Bid or Low Price is the lowest trading price during the day, or the closing bid price on days when the closing price is not available. The field is set to zero if no Bid or Low Price is available.

Table 5.1: Variables in the CRSP stock dataset

by the Office of Management and Budget (OMB), to replace the U.S. Standard Industrial Classification(SIC) system. NAICS is a hierarchical code, containing up to six digits: The first two fields, NAICS sectors, designate general categories of economic activity, the third field, sub-sector, further defines the sector, the fourth field is the industry group, the fifth field is the NAICS industry, and the sixth field represents the national industry (a zero in the 6th digit generally indicates that the NAICS industry and the country industry are the same). According to the 2012 NAICS, businesses related to the U.S. economy are grouped into 20 NAICS sectors, such as utilities, manufacturing, educational services, etc. Each of these sectors includes further classified sub-sectors, the total number of which is 1062.

For each stock, variables recorded include opening price (OPENPRC), closing price (PRC), closing bid (BID), closing ask (ASK), highest trading price (ASKHI), lowest trading price (BIDLO), number of shares sold (VOL), holding period return (RET), and returns without dividends (RETX).

5.2 Empirical Results

In this section, the method is applied to daily stock data during a 10-year period from 2006 to 2015. The primary variables of interest are closing price and trading volume at time t. The candidate covariates are closing ask, closing bid, highest trading price, lowest trading price, holding period return with and without dividends, trading volume, open and close price at time $t-1$ and $t-2$.

Closing price, ask, bid, opening price, highest and lowest trading price are highly correlated, with a correlation greater than 0.999. So further transformation may be meaningful. In the left panel of Figure 5.1, the scatter plot of these variables on log scale clearly demonstrate the strong correlation.

Because the variables we are interested in, i.e., closing price and trading volume, are highly skewed, log-transformation is applied. However, log-transformation is not scale-free, that is, $\log(kx) \neq \log(x)$, and we do not want our model to be affected by changing units, so we take the first difference of the log-transformed closing

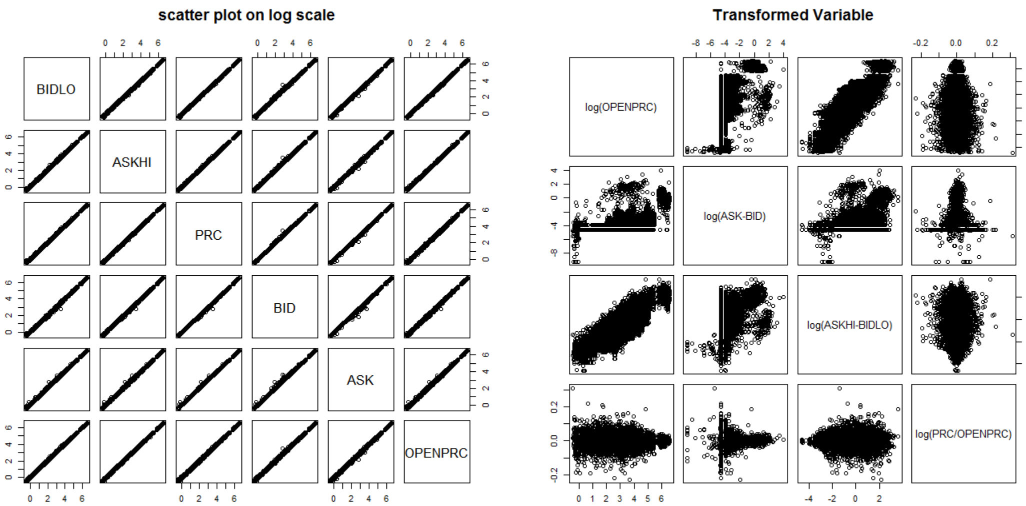

Figure 5.1: Scatter plots of the variables

price and volume. We may make the following transformation

$$\text{LRTRN}(t) = \log(\text{PRC}(t)) - \log(\text{PRC}(t-1)),$$

$$\Delta\text{LVOL}(t) = \log(\text{VOL}(t)+1) - \log(\text{VOL}(t-1)+1).$$

We add 1 to the volume before taking log transformation because the volume could be 0. Thus y_1 is actually log return.

In the CRSP database, the variable SPRTRN is defined as

$$\text{SPRTRN} = \frac{\text{S\&P 500 index at } t}{\text{S\&P 500 index at } t-1} - 1.$$

So we define the log S&P 500 return as

$$\text{LSPRTRN} = \log(\text{SPRTRN} + 1).$$

Further, we construct a variable called SPREAD,

$$\text{SPREAD} = \log(\text{ASKHI}) - \log(\text{BIDLO}).$$

Note that this definition is different from the usual definition of the bid-ask spread, which is defined as ASKHI - BIDLO. In this way, we avoid the strong correlation between closing price, ask price, and bid price, without completely discard the information in ask and bid prices.

To sum up, we construct four variables, namely, log return LRTRN , increase in volume in log scale ΔLVOL, log S&P return LSPRTRN, and SPREAD. We summarize the four variables in Table 5.2 and Figure 5.2.

	LRTRN	ΔLVOL	LSPRTRN	SPREAD
Min	-4.47100	-11.03000	-0.09470	0.00000
Q1	-0.01348	-0.40160	-0.00452	0.02028
Q2	0.00000	-0.01615	0.00072	0.03140
Mean	0.00009	0.00017	0.00019	0.04090
Q3	0.01332	0.37820	0.00580	0.04960
Max.	4.44800	13.91000	0.10960	5.79600

Table 5.2: Summary statistics

We can see from Figure 5.2 that LRTRN, ΔLVOL, and LSPRTRN are pretty symmetric. However, comparing Table 5.2 and Figure 5.2, we can see that all the four variables have long tails. For example, while the central part of the log return is roughly from -0.15 to 0.15, the log return can vary from -4.47 to 4.45. In fact, 0.5% of the log returns are greater than 0.15 or less than -0.15.

Now we specify our model using the variables we just created. Define

$$y_1(t) = \text{LRTRN}(t), \quad y_2(t) = \Delta\text{LVOL}(t),$$

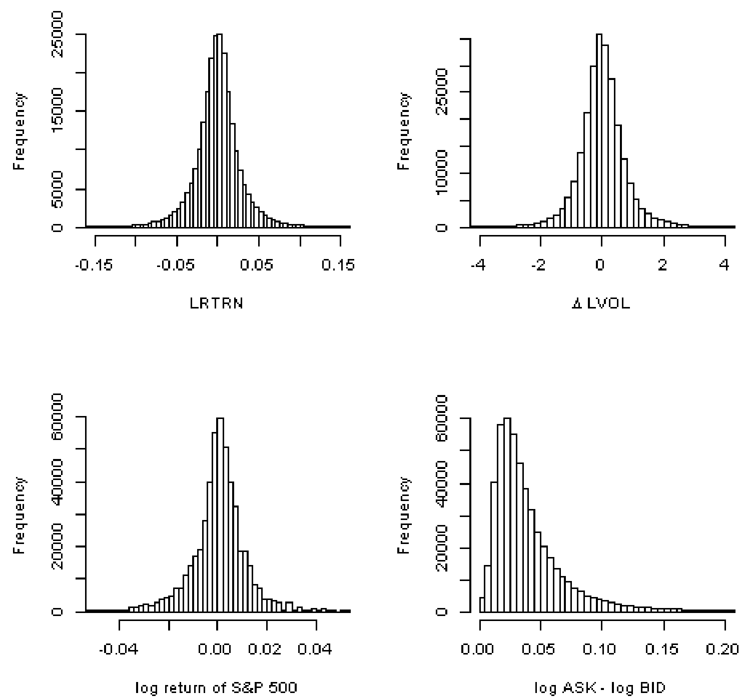

Figure 5.2: Histograms

$$x(t) = \begin{pmatrix} \mathrm{LRTRN}(t-1) \\ \mathrm{LRTRN}(t-2) \\ \Delta\mathrm{LVOL}(t-1) \\ \Delta\mathrm{LVOL}(t-2) \\ \mathrm{LSPRTRN}(t-1) \\ \mathrm{LSPRTRN}(t-2) \end{pmatrix},$$

$$z(t) = (1, \mathrm{SPREAD}(t-1), \mathrm{SPREAD}(t-2))^{\mathsf{T}},$$

then our model can be expressed in the following

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} \alpha_{11}(t) & \alpha_{12}(t) & \alpha_{13}(t) & \alpha_{14}(t) & \alpha_{15}(t) & \alpha_{16}(t) \\ \alpha_{21}(t) & \alpha_{22}(t) & \alpha_{23}(t) & \alpha_{24}(t) & \alpha_{25}(t) & \alpha_{26}(t) \end{pmatrix} x(t)$$
$$+ \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix} z(t) + \begin{pmatrix} \epsilon_{ij1}(t) \\ \epsilon_{ij2}(t) \end{pmatrix}.$$

We do not include an intercept term in $x(t)$ because it is statistically insignificant, which means that there is no significant trend in the mean of log return. An interpretation of this is that any trend in the return should be, say within several days, quickly exploited by the market. Since we will use 10 years of data, and the bandwidth will be much wider than a week, theoretically we should expect that we will not detect any trend in the mean of log return.

At the beginning of the year 2006, there were 2945 stocks traded on the Nasdaq exchange, while at the end of the year 2015, there were 2224 stocks traded on the Nasdaq exchange. However, they were not necessarily the same stocks. There were only 1039 stocks survived from the beginning of 2006 to the end of 2015. These survived stocks belongs to 23 sectors and 292 sub-sectors, according to the NAICS classification code. The distribution of the 1039 stocks among the 23 sectors is documented in Table 5.3.

From the sectors that consist of more than 20 stocks, we randomly choose 5, and then randomly select 20 stocks from these 5 sectors. In total, we include 100 stocks in our analysis. We focus our analysis on 10 years data from 01/03/2006 - 12/31/2015. Thus, for each stock, we have 2,517 observations. In total, we have 251,700 observations, which takes about 64 MB memory space. We use bandwidth

Sector ID	33	52	32	51	54	31	44	55
Stock Count	284	225	101	92	76	31	30	29
Sector ID	45	48	56	42	72	21	53	22
Stock Count	24	22	22	20	20	13	11	10
Sector ID	62	71	23	11	61	81	49	
Stock Count	9	6	5	3	3	2	1	

Table 5.3: Count of stocks by sector.

$h_1 = h_2 = 120$. The algorithm takes approximately five hours to fit the model using R on a 64 bit Windows 8 operating system with Intel Xeon E5 processor, and the peak total memory usage is roughly 60 GB. The memory usage is huge mainly because of matrix manipulation. For example, the dimension of the covariance matrix of the error term is about 251,700 $\times$ 251,700.

The estimated varying coefficients $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are shown in Figure 5.3 and 5.4 respectively. The solid lines represent the estimated varying coefficients, the dashed lines represent the 95% confidence band, and the red horizontal line at 0 is added to make it easier to see if the estimation is significantly different from 0.

For the first response, LRTRN, the coefficients of yesterday (α_{11}) and the day before yesterday's LRTRN (α_{12}) are significantly negative. Further, they share the same pattern. For α_{11}, it is negative and varies between -0.1 and 0 before 2014. During 2014, it becomes very negative, and the minimum is -0.5. α_{12} has a similar pattern, with the magnitude reduce by roughly 50%. The practical meaning is that holding other variables constant, the larger the return in yesterday and the day before yesterday, the more likely that today's return is smaller. In other words, if the stock price increased yesterday, it is more likely that it decreases today; if the stock price also increased the day before yesterday, the chances that it decreases today is even larger, which totally makes sense. This mean reversion relationship is much stronger towards the end of the year 2014. The coefficient of volume is basically positive and barely significant, except in the year 2014. Again, the shape of α_{13} and α_{14} are very similar to each other. A positive coefficient of ΔLVOL means that the more the trading volume increases yesterday, the more likely that the stock price increases today. However, the situation is different in the year 2014, during which the more the trading volume increases yesterday, the more likely that the stock price decreases today. The log return of S&P 500 index has a positive effect on the stock returns, that is, the more the S&P 500 index increases yesterday, the more likely that today's stock price increases.

For the second response, the first difference of log volume, the stock return does not have a significant impact on it. However, the trend of trading volume in previous two days has a significant negative impact on trading volume today. The more the trading volume increased yesterday and the day before yesterday, the more likely that it will decrease today. And the magnitude of yesterday's coefficient is roughly twice the magnitude of the coefficient of the trading volume two days ago. The impact of S&P 500 index on trading volume changes from time to time and goes up and down around 0.

$\hat{\beta}$ is summarized in Table 5.4. The most noticeable pattern is that β_{21} is positive across clusters, ranging from 0.058 to 0.075, and β_{23} is negative across clusters, ranging from -2.25 to -1.59. It suggests that it may be better to include these two covariates in the non-parametric part of the model as well. Moreover, β_{23} is negative

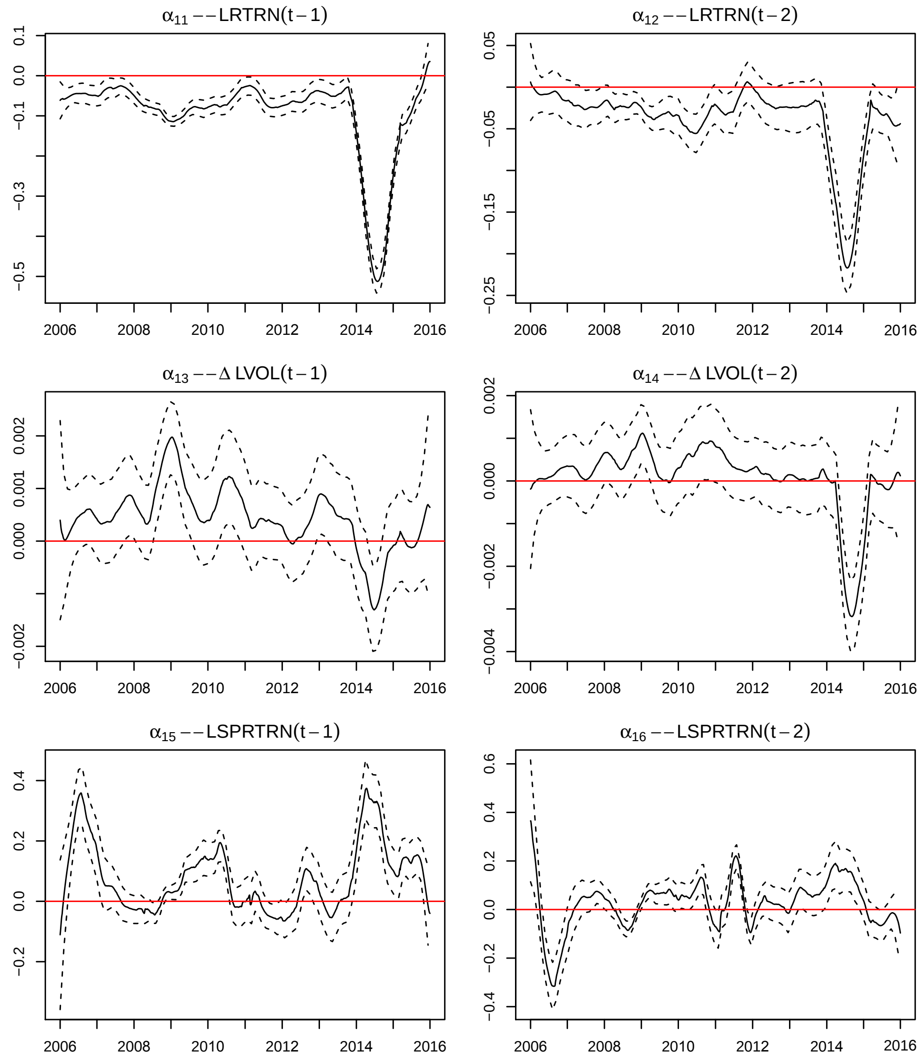

Figure 5.3: Varying coefficients for the first response, LRTRN

means that if in the day before yesterday, the higher the ask-bid spread was, or in other words, the greater the price volatility was, the more likely the trading volume will drop today.

Table 5.5 shows the estimation of the covariance matrix of β. The variances of

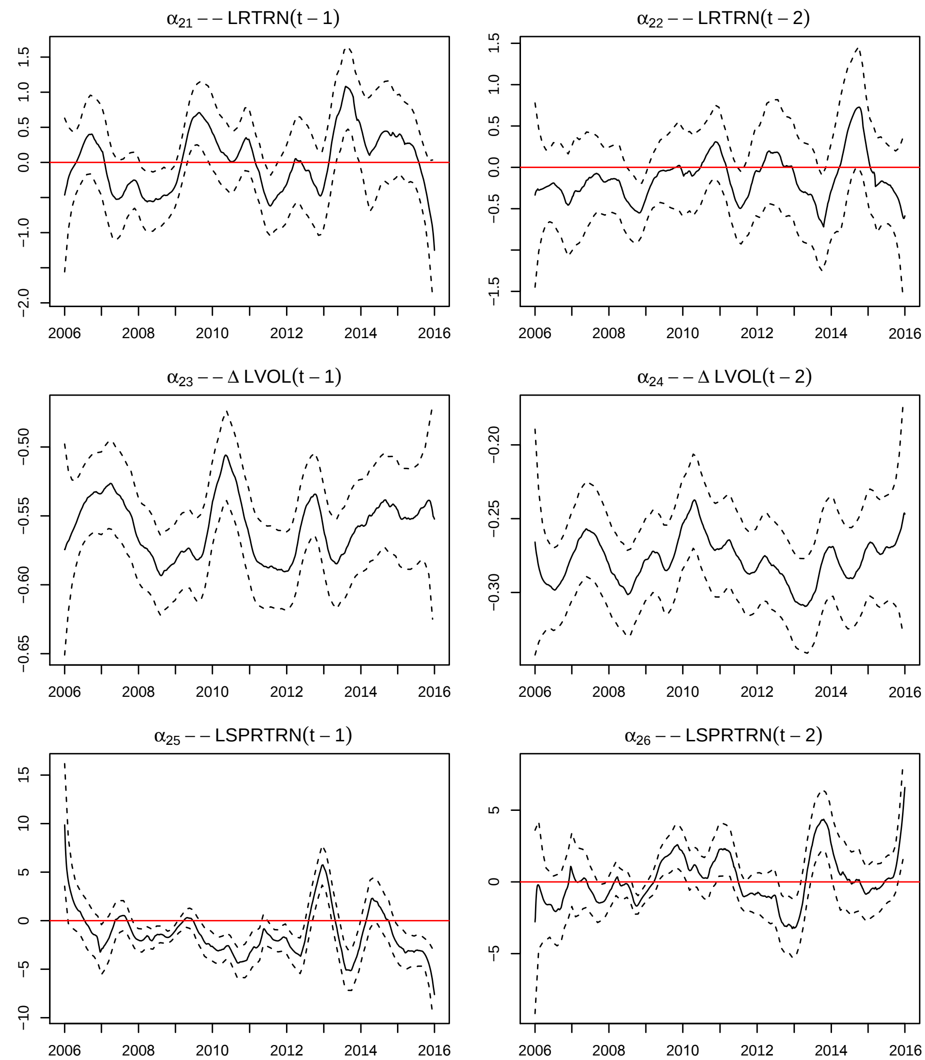

Figure 5.4: Varying coefficients for the second response, ΔLVOL

the random effects for LRTRN, i.e., β_{11}, β_{12}, and β_{13} are much smaller than those for ΔLVOL. However, we can not directly draw the conclusion that the random effect for LRTRN is less important than those for ΔLVOL, because LRTRN and ΔLVOL are of different scales. To facilitate a direct comparison, we calculate the ratio of

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
β_{11}	0.000240	0.001044	0.000480	0.000375	-0.000406
β_{12}	-0.012111	-0.018962	0.001835	0.003878	0.020080
β_{13}	0.012465	0.001550	-0.015010	-0.007859	-0.002974
β_{21}	0.075199	0.108509	0.064889	0.060179	0.057613
β_{22}	0.045929	-0.042046	-0.415582	0.151297	-0.120214
β_{23}	-2.152799	-2.253906	-1.365632	-1.926430	-1.590824

Table 5.4: Predicted β

	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}
β_{11}	0.00000	-0.00001	0.00000	0.00001	0.00000	-0.00010
β_{12}	-0.00001	0.00023	-0.00008	-0.00027	-0.00069	0.00427
β_{13}	0.00000	-0.00008	0.00011	0.00009	0.00125	-0.00298
β_{21}	0.00001	-0.00027	0.00009	0.00043	0.00050	-0.00548
β_{22}	0.00000	-0.00069	0.00125	0.00050	0.04624	-0.06099
β_{23}	-0.00010	0.00427	-0.00298	-0.00548	-0.06099	0.14053

Table 5.5: Estimation of Covariance Matrix of β

	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}
β_{11}	1	-0.84459	-0.01110	0.80022	0.01130	-0.48949
β_{12}	-0.84459	1	-0.49568	-0.84818	-0.21061	0.75017
β_{13}	-0.01110	-0.49568	1	0.41225	0.56498	-0.77149
β_{21}	0.80022	-0.84818	0.41225	1	0.11142	-0.70289
β_{22}	0.01130	-0.21061	0.56498	0.11142	1	-0.75658
β_{23}	-0.48949	0.75017	-0.77149	-0.70289	-0.75658	1

Table 5.6: Estimation of Correlation Matrix of β

the standard deviation of the random effects for LRTRN and ΔLVOL to the third quantile of LRTRN and ΔLVOL, respectively. The ratios are summarized in Table 5.7. We can see that ratios associated with SPREAD(t-1) and SPREAD(t-2) are 1.14, 0.77, 0.57, and 0.99, so they are roughly of the same order. For the random intercept, the ratios are 0.04 and 0.06, which are smaller by at least one order of magnitude.

Recall that the third quantile of SPREAD is 0.0496. Since the ratios for SPREAD in Table 5.7 are around 1, so roughly speaking, one unit percentage change in SPREAD will result in 5% unite percentage change in LRTRN and ΔLVOL.

	LRTRN			ΔLVOL		
SD	0.00052	0.01518	0.01032	0.02081	0.21503	0.37488
Q_3	0.01332	0.01332	0.01332	0.37825	0.37825	0.37825
SD/Q_3	0.03907	1.14025	0.77468	0.05502	0.5685	0.9911

Table 5.7: Ratio of standard deviation to the third quantile.

The estimation of correlation matrix is shown in Table 5.6. Because the number of clusters is limited, the correlation test will fail to reject the null hypothesis that the correlation coefficient is 0.

Chapter 6

Concluding Remarks

6.1 Conclusions

In this thesis, we propose a semiparametric regression model with a hierarchical structure for multivariate longitudinal data. With the design of longitudinal data becoming increasingly complex, our approach can appropriately account for and capture the relationship between responses and their covariates. The longitudinal data, by their very nature, has repeated measurement taken from each subject. And in many applications, subjects are nested within clusters. In addition, when the outcome is continuous, the pattern of change is more reliably characterized by varying-coefficient functions. The incorporation of varying coefficients poses many intricate analytic issues. Finally, each subject can have two outcomes, it is an ideal way to utilize both outcomes in one analysis. These, and many other issues, increase the complexity of longitudinal data analysis. To addresses these issues that arise in analyzing longitudinal data, we proposed multivariate response three-level time-varying mixed-effects models in Chapter 2. This model extends the varying coefficient models, inherits the flexibility of the linear mixed effects models in addressing complex designs and correlation structures, and can use continuous covariates as well as dummy variables in both the fixed or random effects part.

We use the semiparametric regression modeling framework to obtain estimates of the varying-coefficient functions. Our approach is based on the kernel methods of Fan and Gijbels (1996), which uses the local polynomial regression to estimate the varying-coefficient functions. Similar to Fang, Huang and Li (2007), we propose to use the profile likelihood for the estimation of random effects. We have also discussed the properties of local linear regression techniques and the method of profiling and its extension to semiparametric models in Chapter 3. This profile likelihood estimation procedure can be considered better than estimating parameters through iteration, since the latter typically doesn't converge to the overall MLE of the parameters. We also prove that our estimates are consistent and are asymptotically normal. To test whether some or all of the varying coefficient functions are constants, we employ a test based on the comparison of the residual sum of squares (RSS) proposed by Cai, Fan and Yao (2000) for inferences.

The newly proposed semiparametric mixed effects model allows more general design matrices for both the fixed effects and the random effects and includes the univariate model, nonparametric model, and mixed effects model as special cases. The mixed effects and nonparametric model representation allows us to use existing software to fit the models. In our simulation studies of models, we demonstrate that the proposed model has a numerical advantage in various cases.

The application of the multivariate response three-level time-varying mixed-effects model to stock returns and volumes, examined every day for up to ten consecutive years, addresses the nonlinear features in the parameters beyond the capacity of linear time series modeling. It also improves estimation efficiency by incorporating correlation induced by clustering, and modeling stock price and trading volume simultaneously. Karpoff (1987), Chan and Tse (1993) have observed that the relationship between price and volume confirms the usefulness of incorporating volume data to forecast future return. It is interesting that after controlling for other covariates, today's log returns are positively correlated with the change in trading volume of previous two days. The analysis also reveals that S&P 500 index tends to have a positive impact on stock returns, while previous returns have a negative impact on today's return.

6.2 Future Work

For future work, we could explore the following:

Only continuous responses are considered in our model. Longitudinal studies vary in the types of outcomes of interest. The outcomes of interest could be discrete or a bivariate longitudinal outcome of mixed type, more specifically, the combination of a binary with a continuous outcome. Therefore, one can consider broad classes of longitudinal models using link function that may be suitable for analysis.

Second, we have limited our attention in this thesis to error term following a mean-reverting Ornstein-Uhlenbeck process due to its good properties. More broadly, we may consider other covariance structure. The variance of the error term is also assumed to be identical for each subject. However, it may not be the case for the real data. We may add a prior distribution of the variance, such as inverse-gamma to allow a variance that changes from stock to stock.

Third, we would like to adapt our methods to address missing data and model selection problems. The appropriate handling of missing data and an efficient algorithm for the purpose of model selection continue to pose challenges for the analysis of longitudinal data.

Lastly, we employ the cross-validated bandwidth selection in our thesis. It would be interesting to develop a theory for optimal bandwidth choice.

Chapter 7

Appendix

7.1 Derivation of Estimators

Recall

$$y_{dijk} = x_{ijk}^T \alpha_d(t_k) + z_{ijk}^T \beta_{di} + \epsilon_{dijk}.$$

In matrix notation,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} + \begin{pmatrix} Z & \\ & Z \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}.$$

Since α_d is differentiable,

$$\alpha_d(t_{ijk}) \approx \alpha_d(t) + \alpha_d'(t)(t_{ijk} - t).$$

We construct the local linear estimator for $\alpha_d(t)$,

$$\begin{pmatrix} \hat{\alpha}_d(t) \\ \hat{\alpha}_d'(t) \end{pmatrix} = \arg\min_{a,b} \sum_{ijk} \Big(y_{ijk} - x_{ijk}^T(a + b(t_{ijk} - t)) - z_{ijk}^T \beta_{di} \Big)^2 K_{h_d}(t_{ijk} - t),$$

or in matrix notation

$$\begin{pmatrix} \hat{\alpha}_d(t) \\ \hat{\alpha}_d'(t) \end{pmatrix} = \arg\min_{a,b} \left[Y_d - Z\beta_d - (XX_1) \begin{pmatrix} a \\ b \end{pmatrix} \right]^T W \left[Y_d - Z^T\beta_d - (XX_1) \begin{pmatrix} a \\ b \end{pmatrix} \right],$$

where

$$X_1 = \begin{pmatrix} x_{111}^T(t_{111} - t) \\ \vdots \\ x_{n_c n_s n_m}^T(t_{n_c n_s n_m} - t) \end{pmatrix},$$

and W is a weighting matrix as follows,

$$W = \begin{pmatrix} K_{h_d}(t_{111} - t) & & \\ & \ddots & \\ & & K_{h_d}(t_{n_c n_s n_m} - t) \end{pmatrix}.$$

To solve the minimization problem, the first order condition is

$$\begin{pmatrix} X^T \\ X_1^T \end{pmatrix} W \left[Y_d - Z\beta_d - (XX_1) \begin{pmatrix} \hat{\alpha}_d(t) \\ \hat{\alpha}_d'(t) \end{pmatrix} \right] = 0,$$

which implies

$$\begin{pmatrix} \hat{\alpha}_d(t) \\ \hat{\alpha}_d'(t) \end{pmatrix} = \left[\begin{pmatrix} X^T \\ X_1^T \end{pmatrix} W(XX_1) \right]^{-1} \begin{pmatrix} X^T \\ X_1^T \end{pmatrix} W \left[Y_d - Z\beta_d \right].$$

Then, by the formula for matrix inversion in block form,

$$\left[\begin{pmatrix} X^T \\ X_1^T \end{pmatrix} W(XX_1) \right]^{-1} = \begin{pmatrix} X^T W X & X^T W X_1 \\ X_1^T W X & X_1^T W X_1 \end{pmatrix}^{-1}$$

$$=\frac{1}{n_c n_s n_m}\begin{pmatrix} s_{0d} & s_{1d} \\ s_{1d} & s_{2d} \end{pmatrix}^{-1} = \frac{1}{n_c n_s n_m}\begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where

$$A =[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1},$$
$$B =-[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}s_{2d}^{-1}s_{1d}.$$

Hence,

$$\hat{\alpha}_d(t) =\frac{1}{n_c n_s n_m}[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}(X^T - s_{2d}^{-1}s_{1d}X_1^T)W[Y_d - Z\beta_d]$$
$$=\frac{1}{n_c n_s n_m}[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}\sum_{i,j,k}[I_p - s_{1d}s_{2d}^{-1}(t_{ijk} - t)]K_{h_d}(t_{ijk} - t)x_{ijk}(y_{ijk} - z_{ijk}^T\beta_{di}),$$

and $\hat{m}_d$ can be written as

$$\hat{m}_d = S_d(Y_d - Z\beta_d),$$

where each row of S_d is

$$x_{ijk}^T\frac{1}{n_c n_s n_m}[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}(X^T - s_{2d}^{-1}s_{1d}X_1^T)W.$$

Since β_{di} is unobserved, we use its estimator $\hat{\beta}_{di}$,

$$\hat{\alpha}_d(t) =\frac{1}{n_c n_s n_m}[s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}\sum_{i,j,k}[I_p - s_{1d}s_{2d}^{-1}(t_{ijk} - t)]K_{h_d}(t_{ijk} - t)x_{ijk}(y_{ijk} - z_{ijk}^T\hat{\beta}_{di}),$$
$$\hat{m}_d =S_d(Y_d - Z\hat{\beta}_d).$$

Let $L(\beta)$ be the likelihood for β, then

$$-2\log(L(\beta)) \propto(\beta_1^T\beta_2^T)G^{-1}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} +$$

$$+ (Y_1 - m_1 - Z\beta_1)^T R_1^{-1} (Y_1 - m_1 - Z\beta_1)$$
$$+ (Y_2 - m_2 - Z\beta_2)^T R_2^{-1} (Y_2 - m_2 - Z\beta_2).$$

The negative profile log-likelihood of β, up to a constant, is

$$(\beta_1^T \beta_2^T) G^{-1} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$
$$+ (Y_1 - S_d(Y_1 - Z\beta_1) - Z\beta_1)^T R_1^{-1} (Y_1 - S_d(Y_1 - Z\beta_1) - Z\beta_1)$$
$$+ (Y_2 - S_d(Y_2 - Z\beta_2) - Z\beta_2)^T R_2^{-1} (Y_2 - S_d(Y_2 - Z\beta_2) - Z\beta_2),$$

and after simplification is

$$(\beta_1^T \beta_2^T) G^{-1} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$
$$+ (Y_1 - Z\beta_1)^T (I - S_1)^T R_1^{-1} (I - S_1)(Y_1 - Z\beta_1)$$
$$+ (Y_2 - Z\beta_2)^T (I - S_2)^T R_2^{-1} (I - S_2)(Y_2 - Z\beta_2).$$

To maximize the profile likelihood, the first order condition is

$$G^{-1} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - Z^T (I - S_1)^T R_1^{-1} (I - S_1)(Y_1 - Z\beta_1) + Z^T (I - S_2)^T R_2^{-1} (I - S_2)(Y_2 - Z\beta_2) = 0.$$

Solve the first order condition, we get

$$\hat{\beta} = V^{-1} \begin{pmatrix} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Y_1 \\ Z^T (I - S_2)^T R_2^{-1} (I - S_2) Y_2 \end{pmatrix},$$

where

$$V = \begin{pmatrix} Z^T(I - S_1)^T R_1^{-1}(I - S_1)Z & \\ & Z^T(I - S_2)^T R_2^{-1}(I - S_2)Z \end{pmatrix} + G^{-1}.$$

7.2 Glossary

Unless otherwise mentioned, the following definitions of terms and symbols are used:

t	A non-random time point
F	Distribution function of time
f	Probability density function of time
$t_1, \ldots, t_{n_m}$	Independent identically distributed time points
x	Stochastic process with dimension p
z	Stochastic process with dimension q
K	Kernel function
$K_h(t)$	$K(t/h)/h$
μ_r	$\int K(t)t^r\,dt$
ν_r	$\int K^2(t)t^{2r}\,dt$
I_n	$n \times n$ identity matrix
J_n	$n \times n$ matrix of all ones
1_n	$n \times 1$ vector of all ones

$z_{\|\|x}$	$z_{\|\|x}(t) = \mathbf{E}(z(t)x^{\mathsf{T}}(t))[\mathbf{E}x(t)x^{\mathsf{T}}(t)]^{-1}x(t)$
$z_{\perp x}$	$z - z_{\|\|x}$
$C_{x,z}(u,v)$	$\mathbf{E}x(u)z(v)^{\mathsf{T}}$
$\psi_{x,z}(t)$	$\lim_{\substack{u>v \\ u,v\to t}} \frac{\mathbf{E}[x(u)-x(v)][z(u)-z(v)]^{\mathsf{T}}}{u-v}$
$\psi^{d}_{x,z}(t)$	$[\sigma_d^2 \log \frac{1}{\rho_d}]^{-1}\psi_{x,z}(t)$
$u_d(t)$	$-\frac{1}{2}\mu_2 x(t)^{\mathsf{T}}\alpha_d''(t)$
$X_n \xrightarrow{d} Y$	X_n converges to Y in distribution
$X_n \xrightarrow{p} a$	X_n converges to a in probability
$X \sim Y$	X has the same distribution as Y
$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$	$X_1, \ldots, X_n$ are independent identically distributed as X
$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$	$X_1, \ldots, X_n$ are independent identically distributed with distribution function
$X_{(1)}, \ldots, X_{(n)}$	Order statistics of $X_1, \ldots, X_n$

7.3 Theories

7.3.1 Proof of Theorem 3.1

Decompose $\hat{\beta}$ into three components according to $Y_d = m_d + Z\beta_d + E_d$,

$$\begin{aligned}
\hat{\beta} =& C^1 + C^2 + C^3, \\
C^1 =& V^{-1}\begin{pmatrix} Z^T(I-S_1)^T R_1^{-1}(I-S_1)m_1 \\ Z^T(I-S_2)^T R_2^{-1}(I-S_2)m_2 \end{pmatrix}, \\
C^2 =& V^{-1}\begin{pmatrix} Z^T(I-S_1)^T R_1^{-1}(I-S_1)Z\beta_1 \\ Z^T(I-S_2)^T R_2^{-1}(I-S_2)Z\beta_2 \end{pmatrix}, \\
C^3 =& V^{-1}\begin{pmatrix} Z^T(I-S_1)^T R_1^{-1}(I-S_1)E_1 \\ Z^T(I-S_2)^T R_2^{-1}(I-S_2)E_2 \end{pmatrix}.
\end{aligned}$$

Denote

$$C^l = \begin{pmatrix} C^l_1 \\ C^l_2 \end{pmatrix}, \quad C^l_d = \begin{pmatrix} C^l_{d1} \\ \vdots \\ C^l_{di} \\ \vdots \\ C^l_{dn_c} \end{pmatrix},$$

Therefore,

$$\hat{\beta}_{di} = C^1_{di} + C^2_{di} + C^3_{di}.$$

Theorem 7.1. *Under regularity conditions 1-11 (see Section 3.7), there exists a function* g *increasing to infinity, such that as* $n_s, n_m \to \infty$, $n_c^4 = o(n_m)$, $n_m = O(g(n_s))$, $h_d \to 0+$, $h_d {n_m}^{1/2} > \delta$ *for some* $\delta > 0$, $d = 1, 2$, *we have* *(a)* $C^1_{di} = h^2_d [E\psi_{z_{\perp x}, z_{\perp x}}]^{-1} E\psi_{z_{\perp x}, u_d} + o_p(h^2_1) + o_p(h^2_2)$,

(b) $C^2_{di} = \beta_{di} + O_p\left(\frac{1}{n_s n_m}\right)$,

(c) $C^3_{di} = O_p(\frac{1}{n_m})$.

Proof. By Corollary 7.29, there exists a function g increasing to infinity, such that as $n_s, n_m \to \infty$, $n_m = O(g(n_s))$, $n_c = O(n_s)$, $n_c^4 = o(n_m)$, $h_d \to 0+$, $1/h_d = O(\sqrt{n_m})$, $d = 1, 2$, we have,

$$
\begin{aligned}
& n_s n_m V^{-1} \\
&= \left[\frac{1}{n_s n_m} \begin{pmatrix} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Z & \\ & Z^T (I - S_2)^T R_2^{-1} (I - S_2) Z \end{pmatrix} + \frac{1}{n_s n_m} G^{-1} \right]^{-1} \\
&= \begin{pmatrix} I_{n_c} \otimes [E\psi^1_{z,z}]^{-1} & \\ & I_{n_c} \otimes [E\psi^2_{z,z}]^{-1} \end{pmatrix} \\
&\quad + \frac{1}{n_c} \begin{pmatrix} J_{n_c} \otimes ([E\psi^1_{z_{\perp x}, z_{\perp x}}]^{-1} - [E\psi^1_{z,z}]^{-1}) & \\ & J_{n_c} \otimes ([E\psi^2_{z_{\perp x}, z_{\perp x}}]^{-1} - [E\psi^2_{z,z}]^{-1}) \end{pmatrix} \\
&\quad + \frac{1}{n_c} \begin{pmatrix} b_{11} & \cdots & b_{1,2n_c} \\ \vdots & \ddots & \vdots \\ b_{2n_c,1} & \cdots & b_{2n_c,2n_c} \end{pmatrix},
\end{aligned}
$$

where $\forall \epsilon > 0$ elementwisely $\max_{1 \leqslant i,j \leqslant 2n_c} P(|b_{ij}(n_m)| > \epsilon) \to 0$.

(a) Lemma 7.32 shows that

$$
\frac{1}{n_s n_m} Z^T (I - S_d)^T R_d^{-1} (I - S_d) m_d = h_d^2 \begin{pmatrix} E\psi^d_{z_{\perp x}, u_d} \\ \vdots \\ E\psi^d_{z_{\perp x}, u_d} \end{pmatrix} + o_p(h_d^2).
$$

Hence,

$$
\begin{aligned}
C^1 =& V^{-1}\begin{pmatrix} Z^\mathsf{T}(I-S_1)^\mathsf{T}R_1^{-1}(I-S_1)m_1 \\ Z^\mathsf{T}(I-S_2)^\mathsf{T}R_2^{-1}(I-S_2)m_2 \end{pmatrix} \\
=& \begin{pmatrix} h_1^2 1_{n_c} \otimes [\mathbf{E}\psi^1_{z_{\perp x},z_{\perp x}}]^{-1}\mathbf{E}\psi^1_{z_{\perp x},u_1} + o_p(h_1^2) + o_p(h_2^2) \\ h_2^2 1_{n_c} \otimes [\mathbf{E}\psi^2_{z_{\perp x},z_{\perp x}}]^{-1}\mathbf{E}\psi^2_{z_{\perp x},u_2} + o_p(h_1^2) + o_p(h_2^2) \end{pmatrix},
\end{aligned}
$$

where 1_{n_c} is the vector of all-ones with dimension n_c.

(b) Denote $\tilde{G} = \frac{1}{n_s n_m} G$. Let $A = \frac{1}{n_s n_m} V - \tilde{G}$, then,

$$
A = \frac{1}{n_s n_m}\begin{pmatrix} [Z^\mathsf{T}(I-S_1)^\mathsf{T}R_1^{-1}(I-S_1)Z]^{-1} & \\ & [Z^\mathsf{T}(I-S_2)^\mathsf{T}R_2^{-1}(I-S_2)Z]^{-1} \end{pmatrix}.
$$

Using the Woodbury matrix identity (Woodbury 1950),

$$
\begin{aligned}
\left[\frac{1}{n_s n_m} V\right]^{-1} =& A^{-1} - A^{-1}\tilde{G}\left[I + A^{-1}\tilde{G}\right]^{-1} A^{-1} \\
=& A^{-1} - A^{-1}\tilde{G}\left[A + \tilde{G}\right]^{-1} \\
=& A^{-1} - \frac{1}{n_s n_m} A^{-1} G \left[\frac{1}{n_s n_m} V\right]^{-1}.
\end{aligned}
$$

By Corollary 7.29, it can be shown that $A^{-1} G \left[\frac{1}{n_s n_m} V\right]^{-1} = O_p(1)$, therefore,

$$
\begin{aligned}
&\left[\frac{1}{n_s n_m} V\right]^{-1} = A^{-1} + O_p\left(\frac{1}{n_s n_m}\right) \\
=& \begin{pmatrix} \left[\frac{1}{n_s n_m} Z^\mathsf{T}(I-S_1)^\mathsf{T}R_1^{-1}(I-S_1)Z\right]^{-1} & \\ & \left[\frac{1}{n_s n_m} Z^\mathsf{T}(I-S_2)^\mathsf{T}R_2^{-1}(I-S_2)Z\right]^{-1} \end{pmatrix} + O_p\left(\frac{1}{n_s n_m}\right).
\end{aligned}
$$

By Lemma 7.28,

$$\begin{aligned}
&\frac{1}{n_s n_m} Z^T (I - S_d)^T R_d^{-1} (I - S_d) Z \\
&= \begin{pmatrix} E\psi^d_{z,z} + o_p(1) & & \\ & \ddots & \\ & & E\psi^d_{z,z} + o_p(1) \end{pmatrix} \\
&\quad - \frac{1}{n_c} \begin{pmatrix} E\psi^d_{z,z_{||x}} + E\psi^d_{z_{||x},x} - E\psi^d_{z_{||x},z_{||x}} + o_p(1) & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & E\psi^d_{z,z_{||x}} + E\psi^d_{z_{||x},x} - E\psi^d_{z_{||x},z_{||x}} + o_p(1) \end{pmatrix},
\end{aligned}$$

Therefore,

$$\begin{aligned}
C^2 &= \left[\frac{1}{n_s n_m} V\right]^{-1} \begin{pmatrix} \frac{1}{n_s n_m} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Z \beta_1 \\ \frac{1}{n_s n_m} Z^T (I - S_2)^T R_2^{-1} (I - S_2) Z \beta_2 \end{pmatrix} \\
&= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + O_p(\frac{1}{n_s n_m}) \begin{pmatrix} \frac{1}{n_s n_m} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Z \beta_1 \\ \frac{1}{n_s n_m} Z^T (I - S_2)^T R_2^{-1} (I - S_2) Z \beta_2 \end{pmatrix} \\
&= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + O_p\left(\frac{1}{n_s n_m}\right).
\end{aligned}$$

(c) By Lemma 7.34,

$$\frac{1}{n_s n_m} Z^T (I - S_d)^T R_d^{-1} (I - S_d) E_d = O_p(\frac{1}{n_m}).$$

Therefore,

$$C^3 = V^{-1} \begin{pmatrix} Z^T (I - S_1)^T R_1^{-1} (I - S_1) E_1 \\ Z^T (I - S_2)^T R_2^{-1} (I - S_2) E_2 \end{pmatrix} = O_p(\frac{1}{n_m})$$

□

Proof of Theorem 3.1. Under regularity condition 9, $\mathbf{E}\psi^{d}_{z_{\perp x},\epsilon_{d\perp x}} = 0$. By Theorem 7.1,

$$\hat{D} = \frac{1}{n_c}\sum_{i=1}^{n_c}\hat{\beta}_i\hat{\beta}_i^T = \frac{1}{n_c}\sum_{i=1}^{n_c}\beta_i\beta_i^T + o_p(h_1^2) + o_p(h_2^2) + o_p(\frac{1}{n_m}).$$

Hence if $\sqrt{n_c}h_1^2 = O(1)$, $\sqrt{n_c}h_2^2 = O(1)$, $\sqrt{n_c}/n_m = O(1)$, then $\sqrt{n_c}\,\mathrm{vec}(\hat{D})$ and $\sqrt{n_c}\,\mathrm{vec}(\frac{1}{n_c}\sum_{i=1}^{n_c}\beta_i\beta_i^T)$ have the same asymptotic distribution. Applying the central limit theorem to $\sqrt{n_c}\,\mathrm{vec}(\frac{1}{n_c}\sum_{i=1}^{n_c}\beta_i\beta_i^T)$, we complete the proof. □

7.3.2 Proof of Theorem 3.2

For simplicity, write $s_r(t, h_d)$ as s_{rd}. The estimator of $\alpha_d(t)$ is

$$\hat{\alpha}_d(t) = [s_{0d} - s_{1d}s_{2d}^{-1}s_{1d}]^{-1}\frac{1}{n_c n_s n_m}\sum_{i,j,k}[I_p - s_{1d}s_{2d}^{-1}(t - t_{ijk})]K_{h_d}(t_{ijk} - t)x_{ijk}(y_{ijk} - z_{ijk}^T\hat{\beta}_{di}),$$

which can be decomposed into six components,

$$\begin{aligned}
\hat{\alpha}_d(t) =& C_d^1 + C_d^2 + C_d^3 - C_d^4 - C_d^5 - C_d^6,\\
C_d^l =& [s_0(t, h_d) - s_1(t, h_d)s_2^{-1}(t, h_d)s_1(t, h_d)]^{-1}\frac{1}{n_c n_s n_m}M_d^l,\\
M_d^1 =& \tilde{X}(t, h_d)^T m_d, \quad M_d^2 = \tilde{X}^T(t, h_d)Z\beta_d, \quad M_d^3 = \tilde{X}^T(t, h_d)E_d,
\end{aligned}$$

$$\begin{pmatrix} M_1^4 \\ M_2^4 \end{pmatrix} = \begin{pmatrix} \tilde{X}^T(t, h_1)Z & \\ & \tilde{X}^T(t, h_2)Z \end{pmatrix} V^{-1} \begin{pmatrix} Z^T(I - S_1)^T R_1^{-1}(I - S_1)m_1 \\ Z^T(I - S_2)^T R_2^{-1}(I - S_2)m_2 \end{pmatrix},$$

$$\begin{pmatrix} M_1^5 \\ M_2^5 \end{pmatrix} = \begin{pmatrix} \tilde{X}^T(t, h_1)Z & \\ & \tilde{X}^T(t, h_2)Z \end{pmatrix} V^{-1} \begin{pmatrix} Z^T(I - S_1)^T R_1^{-1}(I - S_1)Z\beta_1 \\ Z^T(I - S_2)^T R_2^{-1}(I - S_2)Z\beta_2 \end{pmatrix},$$

$$\begin{pmatrix} M_1^6 \\ M_2^6 \end{pmatrix} = \begin{pmatrix} \tilde{X}^T(t, h_1)Z & \\ & \tilde{X}^T(t, h_2)Z \end{pmatrix} V^{-1} \begin{pmatrix} Z^T(I - S_1)^T R_1^{-1}(I - S_1)E_1 \\ Z^T(I - S_2)^T R_2^{-1}(I - S_2)E_2 \end{pmatrix},$$

where $\tilde{X}$ is defined in the following way. Let

$$\tilde{x}_{ijk}(t,h) = (I_p - s_1(t,h)s_2^{-1}(t,h)(t_k - t))K_h(t_k - t)x_{ijk}.$$

$\tilde{x}_{ijk}(t,h)$ is a p by 1 vector. Now let

$$\tilde{X}_{ij} = \begin{pmatrix} \tilde{x}_{ij1}^T \\ \vdots \\ \tilde{x}_{ijn_m}^T \end{pmatrix}, \quad \tilde{X}_i = \begin{pmatrix} \tilde{X}_{i1} \\ \vdots \\ \tilde{X}_{in_s} \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_{n_c} \end{pmatrix}.$$

Theorem 7.2. *Under regularity conditions 1-11 (see section 3.7), if* t *is an interior point of* $[0,T]$*, then there exists a function* g *increasing to infinity, such that as* $n_s, n_m \to \infty$*,* $n_c^4 = o(n_m)$*,* $n_m = O(g(n_s))$*,* $h_d \to 0+$*,* $h_d {n_m}^{1/2} > \delta$ *for some* $\delta > 0$*,* $d = 1,2$*, we have,*

(a) $C_d^1 = \alpha_d(t) + \frac{1}{2}\mu_2\alpha_d''(t)h_d^2 + o_p(h_d^2)$,

(b) $C_d^4 = h_d^2[Ex(t)x^T(t)]^{-1}[Ex(t)z^T(t)][E\psi_{z_{\perp x},z_{\perp x}}]^{-1}E\psi_{z_{\perp x},u_d} + o_p(h_1^2) + o_p(h_2^2)$,

(c) $C_d^2 - C_d^5 = O_p(\frac{1}{n_s n_m})$,

(d) $C_d^3 = O_p(\frac{1}{\sqrt{n_c n_s}})$,

(e) $C_d^6 = O_p(\frac{1}{\sqrt{n_c n_s n_m}})$.

Proof. (a) By Lemma 7.31, $\forall 0 < \delta < T/2$, as $n_s, n_m \to \infty$, $h_d \to 0+$, $1/h_d = O(\sqrt{n_m})$, elementwisely

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| C_d^1 - \alpha_d(t) - \frac{1}{2}h^2\mu_2\alpha_d''(t) \right| = o_p(h_d^2).$$

(b) This part follows from Corollary 7.19, Lemma 7.21, and Theorem 1.a. By Corollary 7.19, elementwisely

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| [s_0(t,h) - s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1} - \frac{1}{f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1} \right| = o_p(1).$$

By Lemma 7.21,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{n_s n_m} \tilde{X}_i^\top Z_i - f(t)\mathbf{E}x(t)z(t)^\top \right| = o_p(1).$$

By Theorem 7.1.a,

$$V^{-1} \begin{pmatrix} Z^\top (I - S_1)^\top R_1^{-1}(I - S_1)m_1 \\ Z^\top (I - S_2)^\top R_2^{-1}(I - S_2)m_2 \end{pmatrix}$$

$$= \begin{pmatrix} h_1^2 1_{n_c} \otimes [\mathbf{E}\psi^1_{z_{\perp x}, z_{\perp x}}]^{-1}\mathbf{E}\psi^1_{z_{\perp x}, u_1} \\ h_2^2 1_{n_c} \otimes [\mathbf{E}\psi^2_{z_{\perp x}, z_{\perp x}}]^{-1}\mathbf{E}\psi^2_{z_{\perp x}, u_2} \end{pmatrix} + o_p(h_1^2) + o_p(h_2^2).$$

Hence,

$$\begin{pmatrix} C_1^4 \\ C_2^4 \end{pmatrix}$$

$$= \frac{1}{n_c n_s n_m} \begin{pmatrix} [s_{01} - s_{11}s_{21}^{-1}s_{11}]^{-1}\tilde{X}^\top(t,h_1)Z & \\ & [s_{01} - s_{11}s_{21}^{-1}s_{11}]^{-1}\tilde{X}^\top(t,h_2)Z \end{pmatrix} \times$$

$$\times V^{-1} \begin{pmatrix} Z^\top (I - S_1)^\top R_1^{-1}(I - S_1)m_1 \\ Z^\top (I - S_2)^\top R_2^{-1}(I - S_2)m_2 \end{pmatrix}$$

$$= \frac{1}{n_c} \begin{pmatrix} 1_{n_c}^\top \otimes [\mathbf{E}x(t)x(t)^\top]^{-1}\mathbf{E}x(t)z(t)^\top + o_p(1) & \\ & 1_{n_c}^\top \otimes [\mathbf{E}x(t)x(t)^\top]^{-1}\mathbf{E}x(t)z(t)^\top + o_p(1) \end{pmatrix} \times$$

$$\times \begin{pmatrix} h_1^2 1_{n_c} \otimes [\mathbf{E}\psi^1_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi^1_{z_{\perp x}, u_1} + o_p(h_1^2) + o_p(h_2^2) \\ h_2^2 1_{n_c} \otimes [\mathbf{E}\psi^2_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi^2_{z_{\perp x}, u_2} + o_p(h_1^2) + o_p(h_2^2) \end{pmatrix}$$

$$= \begin{pmatrix} h_1^2 [\mathbf{E}x(t)x(t)^\top]^{-1} \mathbf{E}x(t)z(t)^\top [\mathbf{E}\psi^1_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi^1_{z_{\perp x}, u_1} \\ h_2^2 [\mathbf{E}x(t)x(t)^\top]^{-1} \mathbf{E}x(t)z(t)^\top [\mathbf{E}\psi^2_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi^2_{z_{\perp x}, u_2} \end{pmatrix} + o_p(h_1^2) + o_p(h_2^2)$$

$$= \begin{pmatrix} h_1^2 [\mathbf{E}x(t)x(t)^\top]^{-1} \mathbf{E}x(t)z(t)^\top [\mathbf{E}\psi_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi_{z_{\perp x}, u_1} \\ h_2^2 [\mathbf{E}x(t)x(t)^\top]^{-1} \mathbf{E}x(t)z(t)^\top [\mathbf{E}\psi_{z_{\perp x}, z_{\perp x}}]^{-1} \mathbf{E}\psi_{z_{\perp x}, u_2} \end{pmatrix} + o_p(h_1^2) + o_p(h_2^2).$$

(c) By Theorem 1.c,

$$V^{-1} \begin{pmatrix} Z^\top (I - S_1)^\top R_1^{-1} (I - S_1) Z \beta_1 \\ Z^\top (I - S_2)^\top R_2^{-1} (I - S_2) Z \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + O_p\left(\frac{1}{n_s n_m}\right).$$

Therefore,

$$\begin{aligned} C^5 =& \frac{1}{n_c n_s n_m} \begin{pmatrix} [s_{01} - s_{11} s_{21}^{-1} s_{11}]^{-1} \tilde{X}^\top(t, h_1) Z & \\ & [s_{01} - s_{11} s_{21}^{-1} s_{11}]^{-1} \tilde{X}^\top(t, h_2) Z \end{pmatrix} \times \\ & \times V^{-1} \begin{pmatrix} Z^\top (I - S_1)^\top R_1^{-1} (I - S_1) Z \beta_1 \\ Z^\top (I - S_2)^\top R_2^{-1} (I - S_2) Z \beta_2 \end{pmatrix} \\ =& \frac{1}{n_c n_s n_m} \begin{pmatrix} [s_{01} - s_{11} s_{21}^{-1} s_{11}]^{-1} \tilde{X}^\top(t, h_1) Z & \\ & [s_{01} - s_{11} s_{21}^{-1} s_{11}]^{-1} \tilde{X}^\top(t, h_2) Z \end{pmatrix} \times \\ & \times \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + O_p\left(\frac{1}{n_s n_m}\right) \right) \\ =& C^2 + O_p\left(\frac{1}{n_s n_m}\right) \end{aligned}$$

For parts (d) and (e), see the last subsection of Section 7.4. □

Proof of Theorem 3.2. Under the regularity condition 9, we have $\mathbf{E}(x(t)\epsilon_d(t)) = 0$. Therefore by Theorem 7.2,

$$\begin{aligned}\hat{\alpha}_d(t) =& \alpha(t) + \frac{1}{2}\alpha_d''(t)h_d^2 + h_d^2[\mathbf{E}x(t)x^T(t)]^{-1}[\mathbf{E}x(t)z^T(t)][\mathbf{E}\psi_{z_{\perp x},z_{\perp x}}]^{-1}\mathbf{E}\psi_{z_{\perp x},u_d} \\ &+ o_p(h_1^2) + o_p(h_2^2) + O_p(\frac{1}{\sqrt{n_c n_s n_m}}) \\ &+ [s_0(t,h_d) - s_1(t,h_d)s_2^{-1}(t,h_d)s_1(t,h_d)]^{-1}\frac{1}{n_c n_s n_m}\tilde{X}^T(t,h_d)E_d.\end{aligned}$$

By Lemma 7.15, $s_0(t,h_d) = f(t)\mathbf{E}x(t)x(t)^T + o_p(1)$, therefore,

$$\sqrt{n_c n_s}(\hat{\alpha}_d(t) - \alpha_d(t) - \text{bias}) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where $\Sigma = [\mathbf{E}x(t)x(t)^T]^{-1}\mathbf{E}[x(t)x^T(t)\epsilon^2(t)][\mathbf{E}x(t)x(t)^T]^{-1}$. □

7.4 Lemmas

To prove the theorems, we need to directly analyze the asymptotic behaviors of the following terms:

1. $X_i^T R_{dc}^{-1} Z_i$
2. $\tilde{X}(t,h_d)Z$
3. $Z^T(I - S_d)^T R_d^{-1}(I - S_d)Z$
4. $\tilde{X}(t,h_d)m_d$
5. $Z^T(I - S_d)^T R_d^{-1}(I - S_d)m_d$
6. $\tilde{X}(t,h_d)E_d$

7. $Z^T(I - S_d)^T R_d^{-1}(I - S_d)E_d$

We first discuss the asymptotic behaviors of the first two terms, $X^T R_d^{-1} Z$ and $\tilde{X}(t, h_d)Z$, since their asymptotic properties are necessary for the analysis of the third term, $Z^T(I - S_d)^T R_d^{-1}(I - S_d)Z$. Similarly, the asymptotic properties of the fourth term $\tilde{X}(t, h_d)m_d$ is useful to analyze the fifth term, $Z^T(I - S_d)^T R_d^{-1}(I - S_d)m_d$. Altogether, the asymptotic behaviors of the first six terms will help us to prove the theorem about $\hat{\beta}$, i.e., Theorem 1. Along with the asymptotic behaviors of the last term, $\tilde{X}(t, h_d)E_d$, we can prove the theorem about $\hat{\alpha}$, i.e., Theorem 2.

Asymptotic property of $X_i^T R_{dc}^{-1} Z_i$

The following lemma is a generalization of the law of large numbers for random variables to stochastic processes.

Lemma 7.3. *Let* $x_1, x_2, \ldots \overset{i.i.d.}{\sim} x = \{x(t) : 0 \leqslant t \leqslant T\}$*, where* x *has continuous sample paths and* $Ex(t) = 0$*. If* $E \sup_t |x(t)| < \infty$*, then*

$$\sup_{0 \leqslant t \leqslant T} \left| \frac{\sum_{l=1}^{n} x_l(t)}{n} \right| = o_p(1).$$

This lemma can be easily generalized to the situation where t *is in a compact set of a* d *dimensional space* $\mathbb{R}^d$*.*

The lemma is similar to the Glivenko-Cantelli Theorem in Van der Vaart (2000), which also generalizes the law of large numbers. There are important differences. To demonstrate the difference, we reformulate Glivenko-Cantelli Theorem. Let $t_1, t_2, \ldots$ be i.i.d. random variables with distribution function $F(t)$. For $l = 1, 2, \ldots$, define stochastic process

$$x_l(t) = \begin{cases} 0 & t < t_l, \\ 1 & t \geqslant t_l. \end{cases}$$

Then Glivenko-Cantelli Theorem says that

$$\sup_t \left| \frac{\sum_{l=1}^n x_l(t)}{n} - Ex_1(t) \right| = o_p(1),$$

which has the exactly same form in our lemma. However, the difference lies in the restrictions of $x_l(t)$. In Glivenko-Cantelli Theorem, $x_l(t)$ has one point of discontinuity, and is constant elsewhere, while in our lemma, we do not impose restrictions on $x_l(t)$, as long as it is continuous. Another existing result similar to our lemma is Proposition 4 in Mack and Silverman (1982). We cite it here, using our notations to facilitate the representation. Let (t_l, Y_l) be i.i.d. bivariate random variables, and $x_l(t) = K_{h_n}(t_l - t)Y_l$. Note that x_l depends on n.

Theorem 7.4. *[Mack and Silverman (1982, Proposition 4)] If for some $s > 0$, $E|Y_1|^s < \infty$ and $\sup_t \int y^2 f(t, y)\, dt < \infty$, then as $n \to \infty$, $h_n \to 0$, $n^\eta h_n \to \infty$ for some $\eta < 1 - s^{-1}$, we have*

$$\sup_t \left| \frac{\sum_{l=1}^n x_l(t)}{n} - Ex_1(t) \right| = o_p(1).$$

We can see that the difference between Mack and Silverman's proposition and our lemma is that in their proposition, x_l takes a specific form and depends on n. Next, we shall prove our lemma 7.3 by first introducing the following theorem (see Hansen, 2013, Theorem 4.7, or Van der Vaart, 2000, Theorem 18.14).

Theorem 7.5 (Pollard)**.** *Let $x_1, x_2, \ldots$ be stochastic processes on $[0, T]$ with continuous* *sample path. If for every* $\epsilon, \eta > 0$, *there is a grid* $0 = t_1 < \ldots < t_{n'} = T$, *such that*

$$\limsup_{n' \to \infty} P(\max_{1 \leqslant i \leqslant n'-1} \sup_{t_i \leqslant t < t_{i+1}} |x_n(t_i) - x_n(t)| > \epsilon) < \eta,$$

also called Pollard property, and the fidis of x_n *converge weakly to the fidis of* x, *then it holds that* $x_n \xrightarrow{a.s.} x$.

Now, we prove our lemma 7.3.

Proof. Let $\bar{x}_n(t) = \frac{\sum_{l=1}^{n} x_l(t)}{n}$. Because continuous functions from $[0, T]$ to $\mathbb{R}$ form a separable space under uniform metric, to prove that $\bar{x}_n$ weakly converges to a constant process that equals 0 on $[0, T]$, we only need to show that

1. Fidis of $\bar{x}_n$ converge to a vector of zeros;

2. Pollard property: $\forall \epsilon > 0, \delta > 0, \exists\, 0 = t_0 < t_1 < \cdots < t_m = 1$ such that

$$\limsup_{n \to \infty} P(\max_{1 \leqslant i \leqslant m} \sup_{t_{i-1} \leqslant t \leqslant t_i} |\bar{x}_n(t) - \bar{x}_n(t_{i-1})| > \delta) = \epsilon.$$

First, we consider the finite dimensional convergence. It is obvious that $\forall m \in \mathbb{N}$ and $\forall t_k \in [0, T]$, $k = 1, \ldots, m$, as $n \to \infty$, $(\bar{x}_n(t_1), \ldots, \bar{x}_n(t_m)) \xrightarrow{w} (0, \ldots, 0)$.

Second, we consider the tightness. $\forall \delta > 0$, let $t_k = Tk/2^m$, $k = 0, \ldots, 2^m$. Let

$$y_l^m = \max_{1 \leqslant k \leqslant 2^m} \sup_{u, v \in [t_{k-1}, t_k]} |x_l(u) - x_l(v)|.$$

Because $y_l^0 = \sup_{u,v \in [0,T]} |x_l(u) - x_l(v)| \leqslant 2 \sup_{0 \leqslant u \leqslant T} |x_l(u)|$, we have $\infty > \mathbb{E} y_l^0 \geqslant \mathbb{E} y_l^1 \geqslant \mathbb{E} y_l^2 \geqslant \cdots$. We claim that the sequence of expectations declines to 0. Define the set $A_n^m = \{w : y_l^m > \delta\}$, where w is the sample path. Note $A_l^{m+1} \subset A_l^m$. We have $\lim_{m \to \infty} P(A_l^m) = 0$, otherwise, $P(\cap_{m=1}^{\infty} A_l^m) = \lim_{m \to \infty} P(A_l^m) > 0$. Since all paths in $\cap_{m=1}^{\infty} A_l^m$ are discontinuous, with a positive probability, the sample path of

x_l is discontinuous, which contradicts our assumptions. Hence, $\lim_{m\to\infty} Ey_l^m = 0$. $\forall\delta > 0$, find m such that $Ey_l^m < \delta$, then

$$\max_{1\leqslant k\leqslant 2^m} \sup_{u,v\in[t_{k-1},t_k]} |\bar{x}_n(u) - \bar{x}_n(v)| \leqslant \frac{\sum_{l=1}^n y_l^m}{n}$$
$$\Rightarrow P(\max_{1\leqslant k\leqslant 2^m} \sup_{u,v\in[t_{k-1},t_k]} |\bar{x}_n(u) - \bar{x}_n(v)| > \delta) \leqslant P(\frac{\sum_{l=1}^n y_l^m}{n} > \delta).$$

By the law of large numbers, as $n \to \infty$, $P(\frac{\sum_{l=1}^n y_l^m}{n} > \delta) \to 0$, almost surely. Therefore, the tightness of $\bar{x}_1, \bar{x}_2, \ldots$ is established. □

The following lemma says that if a sequence of random variables converges in probability to 0, there exists a sequence of constants increasing to infinity such that after being scaled up by constants, the random variables still converge to 0 in probability.

Lemma 7.6. *Let $X_1, X_2, \ldots,$ be a sequence of random variables. If $X_n = o_p(1)$ as $n \to \infty$, then there exists a function g increasing to infinity such that $X_n = o_p(\frac{1}{g(n)})$.*

Proof. By definition, $\overline{\lim}_{n\to\infty} P(|X_n| > \epsilon) < \delta$, $\forall\epsilon > 0$, $\delta > 0$, or equivalently

$$\underline{\lim}_{n\to\infty} P(|X_n| \leqslant \epsilon) \geqslant \delta, \quad \forall\epsilon > 0, \delta < 1. \tag{$*$}$$

For $\delta < 1$, let the quantile function be defined as $q_n(\delta) = \inf\{\epsilon : P(|X_n| \leqslant \epsilon) \geqslant \delta\}$. Note that a distribution function is right continuous, we have $P(|X_n| \leqslant q_n(\delta)) \geqslant \delta$. Then, $(*)$ is equivalent to $\overline{\lim}_{n\to\infty} q_n(\delta) \leqslant \epsilon$, $\forall\epsilon > 0$, $\delta < 1$, or

$$\lim_{n\to\infty} q_n(\delta) = 0, \quad \forall\delta < 1. \tag{$**$}$$

Let $\delta_1 < \delta_2 < \cdots < 1$ be any sequence such that $\lim_{n\to\infty} \delta_n = 1$. Because $q_n(\delta)$ is increasing with δ, $(**)$ is equivalent to $\lim_{n\to\infty} q_n(\delta_k) = 0$, $k = 1, 2, \ldots$. Similarly,

$X_n = o_p(\frac{1}{g(n)})$ means that

$$\lim_{n\to\infty} g(n)q_n(\delta_k) = 0, \quad k = 1, 2, \ldots. \tag{$***$}$$

Let $n_k = \min\{n : \sup_{n'\geqslant n} q_{n'}(\delta_k) \leqslant 1/2^{2k}\}$, $g(n) = 2^k$ for $n_k \leqslant n < n_{k+1}$. Note that n_k increases with k. It can be shown that $(***)$ holds by examining the limiting behaviors of $g(n)q_n(\delta_k)$ for a fixed k. Suppose $n_{k'} \leqslant n < n_{k'+1}$, and n is sufficient large, such that $k' > k$. Then $q_n(\delta_k) \leqslant q_n(\delta_{k'})$. By definition, $\sup_{n\geqslant n_{k'}} q_n(\delta_{k'}) \leqslant 1/2^{2k'}$, therefore, $q_n(\delta_k) \leqslant q_n(\delta_{k'}) \leqslant 1/2^{2k'}$. Along with $g(n) = 2^{k'}$, we have $g(n)q_n(\delta_k) = 1/2^{k'}$, as $n \to \infty$, $k' \to \infty$ as well, and $(***)$ is established. □

The following lemma is a generalization of the well known result that if Y_1 and Y_2 are i.i.d. $\mathrm{Exp}(\lambda)$, then $Y_1/(Y_1 + Y_2)$ follows uniform distribution according to the relationship between Beta and Gamma distribution. Without loss of generality, we assume $\lambda = 1$. We state but not prove the following result (Ahsanullah, 2015, pp. 12).

Lemma 7.7. *Let* $Y_1, \ldots, Y_{n+1} \overset{i.i.d.}{\sim} \mathrm{Exp}(1)$. *Define* $V_i = \frac{\sum_{j=1}^{i} Y_j}{\sum_{j=1}^{n+1} Y_j}$, $i = 1, \ldots, n$. *Then the random vector* $(V_1, \ldots, V_n)$ *has the same distribution as the ordered statistics of* n *uniformly distributed random variables on* $[0, 1]$.

The following lemma will be useful in obtaining a sequence, $a_1, a_2, \ldots$, such that $X_n = O_p(a_n)$.

Lemma 7.8. *For random variables* $X_1, X_2, \ldots$, *if there is a function* ϕ *such that* $\lim_{\epsilon\to\infty} \phi(\epsilon) = 0$, *and a sequence* $a_1, a_2, \ldots$, *such that* $\forall\epsilon > 0$ *the inequality* $P(|X_n| > \epsilon) < \phi(\epsilon/a_n)$ *holds* *for* $n \geqslant 1$, *then* $X_n = O_p(a_n)$.

Proof. By the definition of big O, we need to prove that $\forall \delta > 0$, there exists $\epsilon > 0$ such that $P(|X_n/a_n| > \epsilon) < \delta, \quad n = 1, 2, \ldots$. By assumptions, $\forall \epsilon > 0$, $P(|X_n/a_n| > \epsilon) < \phi(\epsilon), \quad n = 1, 2, \ldots$. Since ϕ decreases to 0, given $\delta > 0$, there exists $\epsilon > 0$ such that $\phi(\epsilon) < \delta$. Therefore, $P(|X_n/a_n| > \epsilon) < \delta, \quad n = 1, 2, \ldots$. □

Using Schwarz's theorem, if a bivariate function $f(x, y)$ is twice continuously differentiable, $\partial_x(\partial_y f) = \partial_y(\partial_x f) =: \partial_{xy} f$. In fact, it has been shown in his proof that

$$\partial_{xy} f(r, r) = \lim_{x,y \to r} \frac{f(x, x) - f(x, y) - f(y, x) + f(y, y)}{(x - y)^2}.$$

The following lemma considers a situation in which a bivariate function $C(u, v)$ is twice continuously differentiable except on the line $u = v$.

Lemma 7.9. *For a bivariate continuous function* $C(u, v)$ *defined on* $[0, T]^2$, *if it is twice continuously differentiable on* $\{(u, v) : 0 \leqslant u < v \leqslant T\} \cup \{(u, v) : 0 \leqslant v < u \leqslant T\}$ *and all the second partial derivatives are bounded, then*

1. $\lim_{\substack{u<v \\ u,v \to r}} \partial_u C(u, v)$, $\lim_{\substack{u>v \\ u,v \to r}} \partial_u C(u, v)$, $\lim_{\substack{u<v \\ u,v \to r}} \partial_v C(u, v)$, $\lim_{\substack{u>v \\ u,v \to r}} \partial_v C(u, v)$ *exist.*

2. $\lim_{\substack{u>v \\ u,v \to r}} \partial_v C(u, v) - \lim_{\substack{u<v \\ u,v \to r}} \partial_v C(u, v) = \lim_{\substack{u<v \\ u,v \to r}} \partial_u C(u, v) - \lim_{\substack{u>v \\ u,v \to r}} \partial_u C(u, v) =: \psi(r)$, *and* ψ *is continuous on* $[0, T]$.

3. $$\psi(r) = \lim_{\substack{u>v \\ u,v \to r}} \frac{C(u, u) - C(u, v) - C(v, u) + C(v, v)}{u - v}.$$

4. *Define*

$$G(u,v) = \begin{cases} \frac{C(u,u)-C(u,v)-C(v,u)+C(v,v)}{u-v} & u > v \\ \psi(r) & u = v = r \end{cases},$$

then there is a constant L *suth that for* $v \leqslant r \leqslant u$, $|G(u,v) - G(r,r)| \leqslant L(u-v)$.

Proof. In the lower triangle part, $S = \{(u,v) : 0 \leqslant v < u \leqslant T\}$, C is twice continuously differentiable with bounded second order partial derivatives, which implies that $\partial_u C$ and $\partial_v C$ are Lipschitz-continuous on S, that is, there exists a constant M such that for $(u,v) \in S$ and $(u',v') \in S$,

$$\begin{aligned} |\partial_u C(u,v) - \partial_u C(u'v')| \leqslant& M(|u-u'| + |v-v'|), \\ |\partial_v C(u,v) - \partial_v C(u'v')| \leqslant& M(|u-u'| + |v-v'|). \end{aligned}$$

Therefore by Kirszbraun's theorem, $\partial_u C$ and $\partial_v C$ can be continuously extended from S to the closure of S, $\bar{S} = \{(u,v) : 0 \leqslant v \leqslant u \leqslant T\}$, which means that $\lim_{\substack{u>v \\ u,v\to r}} \partial_u C(u,v)$, $\lim_{\substack{u>v \\ u,v\to r}} \partial_v C(u,v)$ exist. Similarly, in the upper triangle part, the limits $\lim_{\substack{u<v \\ u,v\to r}} \partial_u C(u,v)$, $\lim_{\substack{u<v \\ u,v\to r}} \partial_v C(u,v)$ exist. Now for $(u,v) \in S$ define

$$\begin{aligned} f_{u,v}(x) =& C(u,x) - C(v,x), \quad v \leqslant x \leqslant u, \\ G(u,v) =& \frac{f_{u,v}(u) - f_{u,v}(v)}{u-v}. \end{aligned}$$

Since $f_{u,v}$ is defined and continuous on $[v,u]$, and differentiable on (v,u). By the mean value theorem, there exists s lying between x_1 and x_2 such that

$$f_{u,v}(x_1) - f_{u,v}(x_2) = f'_{u,v}(s)(x_1 - x_2).$$

Let $x_1 = u$ and $x_2 = v$, we have

$$G(u,v) = f'_{u,v}(s) = \partial_v C(u,s) - \partial_v C(v,s).$$

Note that $\partial_v C(x,s)$ is a function of x, which is not necessarily well defined on $[v,u]$, due to the potential non-existence of $\partial_v C(s,s)$. We can not apply the mean value theorem to $\partial_v C(u,s) - \partial_v C(v,s)$ with respect to the first argument. G can be continuously extended to $\bar{S}$ by defining

$$\begin{aligned} G(r,r) &= \lim_{\substack{u>v \\ u,v\to r}} \frac{C(u,u) - C(v,u) - C(u,v) + C(v,v)}{u-v} \\ &= \lim_{\substack{u>v \\ u,v\to r}} (\partial_v C(u,s) - \partial_v C(v,s)) \\ &= \lim_{\substack{u>v \\ u,v\to r}} \partial_v C(u,v) - \lim_{\substack{u<v \\ u,v\to r}} \partial_v C(u,v). \end{aligned}$$

Similarly, by considering $g_{u,v}(x) = C(x,u) - C(x,v)$, we can also define

$$\begin{aligned} G(r,r) &= \lim_{\substack{u>v \\ u,v\to r}} \frac{C(u,u) - C(u,v) - C(v,u) + C(v,v)}{u-v} \\ &= \lim_{\substack{u>v \\ u,v\to r}} \frac{g_{u,v}(u) - g_{u,v}(v)}{u-v} = \lim_{\substack{u>v \\ u,v\to r}} g'_{u,v}(s) \\ &= \lim_{\substack{u>v \\ u,v\to r}} (\partial_u C(s,u) - \partial_u C(s,v)) \\ &= \lim_{\substack{u<v \\ u,v\to r}} \partial_u C(u,v) - \lim_{\substack{u>v \\ u,v\to r}} \partial_u C(u,v). \end{aligned}$$

By the uniqueness of limit, we conclude that

$$\lim_{\substack{u>v \\ u,v\to r}} \partial_v C(u,v) - \lim_{\substack{u<v \\ u,v\to r}} \partial_v C(u,v) = \lim_{\substack{u<v \\ u,v\to r}} \partial_u C(u,v) - \lim_{\substack{u>v \\ u,v\to r}} \partial_u C(u,v) =: \psi(r).$$

Now, we prove that for $v \leqslant r \leqslant u$, there is a constant L such that $|G(u,v) - G(r,r)| \leqslant L(u-v)$. The case in which $u = v$ is obvious. Consider the case $u \neq v$. Let $v' > u'$

such that $v < v'$and $u' < u$, then

$$\begin{aligned}
&|G(u,v) - G(u',v')| \\
=&|(\partial_v C(u,s) - \partial_v C(v,s)) - (\partial_v C(u',s') - \partial_v C(v',s'))| \\
=&|(\partial_v C(u,s) - \partial_v C(u',s')) - (\partial_v C(v,s) - \partial_v C(v',s'))| \\
\leqslant&|\partial_v C(u,s) - \partial_v C(u',s')| + |\partial_v C(v,s) - \partial_v C(v',s')| \\
\leqslant&M(|u-u'| + |s-s'|) + M(|v-v'| + |s-s'|) \\
=&M(|u-u'| + |v-v'|) + 2M|s-s'| \\
\leqslant&M(|u-u'| + |v-v'|) + 2M|\max\{u,u'\} - \min\{v,v'\}| \\
\leqslant&M(|u-u'| + |v-v'|) + 2M(|u-v| + |u-v'| + |u'-v| + |u'-v'|).
\end{aligned}$$

Therefore by continuity,

$$\begin{aligned}
&|G(u,v) - G(r,r)| = \lim_{\substack{u'>v' \\ u',v'\to r}} |G(u,v) - G(u',v')| \\
\leqslant& \lim_{\substack{u'>v' \\ u',v'\to r}} \left[M(|u-u'| + |v-v'|) + 2M(|u-v| + |u-v'| + |u'-v| + |u'-v'|)\right] \\
=&M(|u-r| + |v-r|) + 2M(|u-v| + |u-r| + |r-v| + 0) \\
\leqslant&M(|u-v| + |v-u|) + 2M(|u-v| + |u-v| + |u-v|) \leqslant 8M(u-v).
\end{aligned}$$

□

Lemma 7.10. *Let* $(x_1, z_1), \dots, (x_{n_s}, z_{n_s}) \overset{i.i.d.}{\sim} (x, z)$, *where* $(x, z) = \{(x(t), z(t)) : 0 \leqslant t \leqslant T\}$ *is a continuous stochastic process,* $x(t) = (x^1(t), \dots, x^p(t))^\mathsf{T}$ *is* $p \times 1$, *and* $z(t) =$

$(z^1(t), \ldots, z^q(t))^\mathsf{T}$ *is* $q \times 1$. *Let* $t_1, \ldots, t_{n_m} \overset{i.i.d.}{\sim} F$, *where* F *has bounded derivatives. Let*

$$R = \begin{pmatrix} \sigma^2 & \rho^{t_{(2)}-t_{(1)}}\sigma^2 & \rho^{t_{(3)}-t_{(1)}}\sigma^2 & \cdots & \rho^{t_{(n_m)}-t_{(1)}}\sigma^2 \\ \rho^{t_{(2)}-t_{(1)}}\sigma^2 & \sigma^2 & \rho^{t_{(3)}-t_{(2)}}\sigma^2 & \cdots & \rho^{t_{(n_m)}-t_{(2)}}\sigma^2 \\ \rho^{t_{(3)}-t_{(1)}}\sigma^2 & \rho^{t_{(2)}-t_{(1)}}\sigma^2 & \sigma^2 & \cdots & \rho^{t_{(n_m)}-t_{(3)}}\sigma^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t_{(n_m)}-t_{(1)}}\sigma^2 & \rho^{t_{(n_m)}-t_{(2)}}\sigma^2 & \rho^{t_{(n_m)}-t_{(3)}}\sigma^2 & \cdots & \sigma^2 \end{pmatrix},$$

where $0 < \rho < 1$. *Let* $x_{jk} = x_j(t_k)$, $z_{jk} = z_j(t_k)$,

$$X_j = \begin{pmatrix} x_{j1}^\mathsf{T} \\ \vdots \\ x_{jn_m}^\mathsf{T} \end{pmatrix}, \quad Z_j = \begin{pmatrix} z_{j1}^\mathsf{T} \\ \vdots \\ z_{jn_m}^\mathsf{T} \end{pmatrix},$$

and let $C_{x,z}(u, v) = \mathbb{E}x(u)z^\mathsf{T}(v)$. *If elementwisely,* $\mathbb{E}\sup_{0 \leqslant t \leqslant T}[x(t)]^2 < \infty$, $\mathbb{E}\sup_{0 \leqslant t \leqslant T}[z(t)]^2 < \infty$, $C_{x,z}$ *is continuous on* $[0, T]^2$, *and has bounded continuous second partial derivatives on* $\{(u, v) : 0 \leqslant u < v \leqslant T\} \cup \{(u, v) : 0 \leqslant v < u \leqslant T\}$, *then there exists a function* g *increasing to infinity such that as* $n_s \to \infty$ *and* $n_m = O(g(n_s))$, *we have*

$$\frac{1}{n_s}\sum_{j=1}^{n_s} X_j^\mathsf{T} R^{-1} Z_j = [-2\sigma^2 \log \rho]^{-1} \sum_{k=1}^{n_m} \psi_{x,z}(t_k) + O_p(1),$$

where $\psi_{x,z}(t) = \lim_{\substack{u<v \\ u,v \to t}} \partial_u C_{x,z} - \lim_{\substack{u>v \\ u,v \to t}} \partial_u C_{x,z} = \lim_{\substack{u>v \\ u,v \to t}} \partial_v C_{x,z} - \lim_{\substack{u<v \\ u,v \to t}} \partial_v C_{x,z}$.

Proof. We prove the case in which $p = q = 1$. The multivariate case is similar. In addition, to simplify the notation, we use t_i to represent $t_{(i)}$ in this proof.

$$R = \begin{pmatrix} \sigma^2 & \rho^{t_2-t_1}\sigma^2 & \rho^{t_3-t_1}\sigma^2 & \cdots & \rho^{t_{n_m}-t_1}\sigma^2 \\ \rho^{t_2-t_1}\sigma^2 & \sigma^2 & \rho^{t_3-t_2}\sigma^2 & \cdots & \rho^{t_{n_m}-t_2}\sigma^2 \\ \rho^{t_3-t_1}\sigma^2 & \rho^{t_2-t_1}\sigma^2 & \sigma^2 & \cdots & \rho^{t_{n_m}-t_3}\sigma^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t_{n_m}-t_1}\sigma^2 & \rho^{t_{n_m}-t_2}\sigma^2 & \rho^{t_{n_m}-t_3}\sigma^2 & \cdots & \sigma^2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & & \\ \rho^{t_2-t_1} & 1 & & & \\ \rho^{t_3-t_1} & \rho^{t_3-t_2} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^{t_{n_m}-t_1} & \rho^{t_{n_m}-t_2} & \rho^{t_{n_m}-t_3} & \cdots & 1 \end{pmatrix} \times$$

$$\sigma^2 \begin{pmatrix} 1 & & & & \\ & 1-\rho^{2(t_2-t_1)} & & & \\ & & 1-\rho^{2(t_3-t_2)} & & \\ & & & \ddots & \\ & & & & 1-\rho^{2(t_{n_m}-t_{n_m-1})} \end{pmatrix} \times$$

$$\begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \cdots & \rho^{t_{n_m}-t_1} \\ & 1 & \rho^{t_3-t_2} & \cdots & \rho^{t_{n_m}-t_2} \\ & & 1 & \cdots & \rho^{t_{n_m}-t_3} \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}.$$

$$R^{-1} = \begin{pmatrix} 1 & -\rho^{t_2-t_1} & & & \\ & 1 & -\rho^{t_3-t_2} & & \\ & & 1 & \ddots & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} \times \sigma^{-2} \begin{pmatrix} 1 & & & & \\ & \frac{1}{1-\rho^{2(t_2-t_1)}} & & & \\ & & \frac{1}{1-\rho^{2(t_3-t_2)}} & & \\ & & & \ddots & \\ & & & & \frac{1}{1-\rho^{2(t_{n_m}-t_{n_m-1})}} \end{pmatrix} \times$$

$$\begin{pmatrix} 1 & & & & \\ -\rho^{t_2-t_1} & 1 & & & \\ & -\rho^{t_3-t_2} & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \end{pmatrix}.$$

$$
\begin{aligned}
& X_j^T R^{-1} Z_j \\
=& \sigma^{-2} x_{j1} z_{j1} + \sigma^{-2} \sum_{k=2}^{n_m} \frac{1}{1-\rho^{2(t_k - t_{k-1})}} (x_{jk} - \rho^{t_k - t_{k-1}} x_{j,k-1})(z_{jk} - \rho^{t_k - t_{k-1}} z_{j,k-1}) \\
=& \sigma^{-2} x_{j1} z_{j1} + \sigma^{-2} \sum_{k=2}^{n_m} \frac{1}{1-\rho^{2(t_k - t_{k-1})}} [(1-\rho^{t_k - t_{k-1}}) x_{jk} + \rho^{t_k - t_{k-1}} (x_{j,k} - x_{j,k-1})] \times \\
& [(1-\rho^{t_k - t_{k-1}}) z_{jk} + \rho^{t_k - t_{k-1}} (z_{j,k} - z_{j,k-1})] \\
=& \sigma^{-2} x_{j1} z_{j1} + \sigma^{-2} \sum_{k=2}^{n_m} \frac{1-\rho^{t_k - t_{k-1}}}{1+\rho^{t_k - t_{k-1}}} x_{jk} z_{jk} \qquad (1) \\
& + \sigma^{-2} \sum_{k=2}^{n_m} \frac{\rho^{t_k - t_{k-1}}}{1+\rho^{t_k - t_{k-1}}} (x_{jk}(z_{jk} - z_{j,k-1}) + (x_{jk} - x_{j,k-1}) z_{jk}) \qquad (2) \\
& + \sigma^{-2} \sum_{k=2}^{n_m} \frac{\rho^{2(t_k - t_{k-1})}}{1-\rho^{2(t_k - t_{k-1})}} (x_{jk} - x_{j,k-1})(z_{jk} - z_{j,k-1}). \qquad (3)
\end{aligned}
$$

(1)

$$
\begin{aligned}
& \left| \frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} \frac{1-\rho^{t_k - t_{k-1}}}{1+\rho^{t_k - t_{k-1}}} x_{jk} z_{jk} \right| \\
\leqslant & \frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} \left| \frac{1-\rho^{t_k - t_{k-1}}}{1+\rho^{t_k - t_{k-1}}} x_{jk} z_{jk} \right| \\
\leqslant & \frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} (1-\rho^{t_k - t_{k-1}}) |x_{jk} z_{jk}| \\
\leqslant & \frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} (1-\rho^{t_k - t_{k-1}}) \frac{x_{jk}^2 + z_{jk}^2}{2} \\
\leqslant & \frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} (1-\rho^{t_k - t_{k-1}}) \frac{\sup_t x_j^2(t) + \sup_t z_j^2(t)}{2} \\
= & \sum_{k=2}^{n_m} (1-\rho^{t_k - t_{k-1}}) \sum_{j=1}^{n_s} \frac{\sup_t x_j^2(t) + \sup_t z_j^2(t)}{2 n_s}.
\end{aligned}
$$

Because

$$\frac{\sum_{j=1}^{n_s}\sup_t x_j^2(t)}{n_s} = O_p(1), \quad \frac{\sum_{j=1}^{n_s}\sup_t z_j^2(t)}{n_s} = O_p(1),$$
$$\sum_{k=2}^{n_m}(1-\rho^{t_k-t_{k-1}}) \leqslant \sum_{k=2}^{n_m} -\log\rho(t_k-t_{k-1}) \leqslant -T\log\rho,$$

we have

$$\frac{1}{n_s}\sum_{k=2,j=1}^{n_m,n_s}\frac{1-\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}x_{jk}z_{jk} = O_p(1).$$

(2)

$$\frac{1}{n_s}\sum_{k=2,j=1}^{n_m,n_s}\frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}(x_{jk}-x_{j,k-1})z_{jk} = \sum_{k=2}^{n_m}\left(\frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}\frac{\sum_{j=1}^{n_s}(x_{jk}-x_{j,k-1})z_{jk}}{n_s}\right)$$

$$= \sum_{k=2}^{n_m}\frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}\left(\frac{\sum_{j=1}^{n_s}(x_{jk}-x_{j,k-1})z_{jk}}{n_s} - \mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k]\right) \tag{2.1}$$

$$+ \sum_{k=2}^{n_m}\frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}\mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k]. \tag{2.2}$$

According to Lemma 7.3, as $n_s \to \infty$,

$$\sup_{0\leqslant t'\leqslant t\leqslant T}\left|\frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))z_j(t)}{n_s} - \mathbf{E}(x(t)-x(t'))z(t)\right| = o_p(1).$$

Lemma 7.6 guarantees that there exists a function g_1 increasing to infinity such that as $n_s \to \infty$,

$$\sup_{0\leqslant t'\leqslant t\leqslant T}\left|\frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))z_j(t)}{n_s} - \mathbf{E}(x(t)-x(t'))z(t)\right| = o_p(\frac{1}{g_1(n_s)}).$$

Hence,

$$\left|\sum_{k=2}^{n_m}\frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}}\left(\frac{\sum_{j=1}^{n_s}(x_{jk}-x_{j,k-1})z_{jk}}{n_s} - \mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k]\right)\right|$$

$$
\begin{aligned}
&\leqslant \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}} \left| \frac{\sum_{j=1}^{n_s}(x_{jk}-x_{j,k-1})z_{jk}}{n_s} - \mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k] \right| \\
&\leqslant \sum_{k=2}^{n_m} \left| \frac{\sum_{j=1}^{n_s}(x_{jk}-x_{j,k-1})z_{jk}}{n_s} - \mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k] \right| \\
&\leqslant \sum_{k=2}^{n_m} \sup_{0\leqslant t'\leqslant t\leqslant T} \left| \frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))z_j(t)}{n_s} - \mathbf{E}[(x(t)-x(t'))z(t)] \right| \\
&\leqslant n_m \sup_{0\leqslant t'\leqslant t\leqslant T} \left| \frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))z_j(t)}{n_s} - \mathbf{E}[(x(t)-x(t'))z(t)] \right| = o_p(\frac{n_m}{g_1(n_s)}).
\end{aligned}
$$

That is, as $n_s \to \infty$, (2.1) $= o_p(\frac{n_m}{g_1(n_s)})$.

Now, we analyze (2.2). Since $\partial_u C_{x,z}$ is bounded, suppose $\partial_u C_{x,z} \leqslant C$,

$$
\begin{aligned}
&\left| \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}}{1+\rho^{t_k-t_{k-1}}} \mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k] \right| \\
=&\left| \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}(t_k-t_{k-1})}{1+\rho^{t_k-t_{k-1}}} \frac{\mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k]}{t_k-t_{k-1}} \right| \\
\leqslant& \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}(t_k-t_{k-1})}{1+\rho^{t_k-t_{k-1}}} \left| \frac{\mathbf{E}[(x(t_k)-x(t_{k-1}))z(t_k)|t_{k-1},t_k]}{t_k-t_{k-1}} \right| \\
=&\text{ãŁŁ} \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}(t_k-t_{k-1})}{1+\rho^{t_k-t_{k-1}}} \left| \frac{C_{xz}(t_k,t_k)-C_{xz}(t_{k-1},t_k)}{t_k-t_{k-1}} \right| \\
=& \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}(t_k-t_{k-1})}{1+\rho^{t_k-t_{k-1}}} \left| \partial_u C_{x,z}(\tilde{t}_k,t_k) \right| \\
\leqslant& \sum_{k=2}^{n_m} \frac{\rho^{t_k-t_{k-1}}(t_k-t_{k-1})}{1+\rho^{t_k-t_{k-1}}} C \leqslant C \sum_{k=2}^{n_m}(t_k-t_{k-1}) \leqslant CT.
\end{aligned}
$$

That is, (2.2) $= O_p(1)$.

(3) Now consider

$$
\frac{1}{n_s} \sum_{k=2,j=1}^{n_m,n_s} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} (x_{jk}-x_{j,k-1})(z_{jk}-z_{j,k-1})
$$

$$=\sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \left(\frac{\sum_{j=1}^{n_s}(x_j(t_k)-x_j(t_{k-1}))(z_j(t_k)-z_j(t_{k-1}))}{n_s} - \right.$$

$$\left. \mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k] \right) \tag{3.1}$$

$$+\sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]. \tag{3.2}$$

According to Lemma 7.3, as $n_s \to \infty$,

$$\sup_{0\leqslant t'\leqslant t\leqslant T} \left| \frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))(z_j(t)-z_j(t'))}{n_s} - \mathbf{E}[(x(t)-x(t_{k'}))(z_j(t)-z_j(t'))] \right| = o_p(1).$$

According to Lemma 7.6, there exists a function g_2 increasing to infinity such that as $n_s \to \infty$

$$\sup_{0\leqslant t'\leqslant t\leqslant T} \left| \frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))(z_j(t)-z_j(t'))}{n_s} - \mathbf{E}[(x(t)-x(t_{k'}))(z_j(t)-z_j(t'))] \right| = o_p(\frac{1}{g_2(n_s)})$$

Hence,

$$\left| \sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \left(\frac{\sum_{j=1}^{n_s}(x_j(t_k)-x_j(t_{k-1}))(z_j(t_k)-z_j(t_{k-1}))}{n_s} - \right. \right.$$

$$\left. \left. \mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k] \right) \right|$$

$$\leqslant \sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \left| \frac{\sum_{j=1}^{n_s}(x_j(t_k)-x_j(t_{k-1}))(z_j(t_k)-z_j(t_{k-1}))}{n_s} - \right.$$

$$\left. \mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k] \right|$$

$$\leqslant \sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \sup_{0\leqslant t'\leqslant t\leqslant T} \left| \frac{\sum_{j=1}^{n_s}(x_j(t)-x_j(t'))(z_j(t)-z_j(t'))}{n_s} - \right.$$

$$\left. - \mathbf{E}[(x(t)-x(t_{k'}))(z_j(t)-z_j(t'))] \right|$$

$$=o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}}=o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{\rho^{2(t_k-t_{k-1})}}{1+\rho^{t_k-t_{k-1}}}\frac{1}{1-\rho^{t_k-t_{k-1}}}$$
$$\leqslant o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{1}{1-\rho^{t_k-t_{k-1}}}=o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{1}{F(t_k)-F(t_{k-1})}\frac{F(t_k)-F(t_{k-1})}{1-\rho^{t_k-t_{k-1}}}.$$

We claim that there exists a constant, say C, such that $\sup_{0\leqslant t'<t\leqslant T}\frac{F(t)-F(t')}{1-\rho^{t-t'}}\leqslant C$. Therefore,

$$\begin{aligned}
&o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{1}{F(t_k)-F(t_{k-1})}\frac{F(t_k)-F(t_{k-1})}{1-\rho^{t_k-t_{k-1}}}\\
\leqslant&o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{1}{F(t_k)-F(t_{k-1})}C\\
=&o_p(\frac{1}{g_2(n_s)})\sum_{k=2}^{n_m}\frac{1}{F(t_k)-F(t_{k-1})}.
\end{aligned}$$

By Lemma 7.7, there exist $Y_1,\dots,Y_{n_m+1}\overset{\text{i.i.d.}}{\sim}\mathrm{Exp}(1)$, such that $F(t_k)-F(t_{k-1})=\frac{Y_k}{\sum_{i=1}^{n_m+1}Y_i}$, for $k=2,\dots,n_m$. So

$$\sum_{k=2}^{n_m}\frac{1}{F(t_k)-F(t_{k-1})}=\sum_{k=2}^{n_m}\frac{\sum_{i=1}^{n_m+1}Y_i}{Y_k}=\sum_{i=1}^{n_m+1}Y_i\sum_{k=2}^{n_m}\frac{1}{Y_k}$$

By the central limit theorem, $\sum_{i=1}^{n_m+1}Y_i=O_p(\sqrt{n_m})$, and further we claim that $\sum_{k=2}^{n_m}\frac{1}{Y_k}=O_p(n_m)$. Therefore,

$$\sum_{i=1}^{n_m+1}Y_i\sum_{k=2}^{n_m}\frac{1}{Y_k}=O_P(n_m^{1.5}).$$

Hence, we conclude that

$$\left|\sum_{k=2}^{n_m}\frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}}\left(\frac{\sum_{j=1}^{n_s}(x_j(t_k)-x_j(t_{k-1}))(z_j(t_k)-z_j(t_{k-1}))}{n_s}-\right.\right.$$
$$\left.\left.\mathbb{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]\right)\right|=o_p(\frac{n_m^{1.5}}{g_2(n_s)}).$$

That is, as $n_s \to \infty$, (3.1) $= o_p(\frac{n_m^{1.5}}{g_2(n_s)})$. Now we prove the claim that there is a constant C such that

$$\sup_{0\leqslant t'<t\leqslant T} \frac{F(t)-F(t')}{1-\rho^{t-t'}} \leqslant C.$$

There exist u and v that lie between t' and t such that

$$F(t)-F(t') = F'(u)(t-t'), \quad \rho^{t'}-\rho^{t} = (\log\rho)\rho^{v}(t'-t).$$

Hence,

$$\frac{F(t)-F(t')}{1-\rho^{t-t'}} = \frac{F(t)-F(t')}{\rho^{-t'}(\rho^{t'}-\rho^{t})} = \frac{F'(u)(t-t')}{\rho^{-t'}(\log\rho)\rho^{v}(t'-t)} = \frac{F'(u)}{(-\log\rho)\rho^{v-t'}} \leqslant \frac{F'(u)}{(-\log\rho)\rho^{T}}.$$

Since F' is bounded, we have

$$\sup_{0\leqslant t'<t\leqslant T} \frac{F(t)-F(t')}{1-\rho^{t-t'}} \leqslant \frac{\sup_t F'(t)}{(-\log\rho)\rho^{T}}.$$

Next we prove the claim that $\sum_{k=2}^{n_m}\frac{1}{Y_k} = O_p(n_m)$. Let $\phi(x) = 1/x$. Note that ϕ is convex. $\forall \epsilon > 0$

$$\begin{aligned} P(\sum_{k=2}^{n_m}\frac{1}{Y_k} > \epsilon) =& P(\frac{1}{n_m-1}\sum_{k=2}^{n_m}\phi(Y_k) > \frac{\epsilon}{n_m-1}) \\ \leqslant& P(\phi(\frac{1}{n_m-1}\sum_{k=2}^{n_m}Y_k) > \frac{\epsilon}{n_m-1}) \\ \leqslant& \frac{\mathbf{E}\phi(\frac{1}{n_m-1}\sum_{k=2}^{n_m}Y_k)}{\frac{\epsilon}{n_m-1}} = \frac{(n_m-1)^2\mathbf{E}\frac{1}{\sum_{k=2}^{n_m}Y_k}}{\epsilon}. \end{aligned}$$

Since $Y_k \sim \text{Exp}(1) = \Gamma(1,1)$, so $\sum_{k=2}^{n_m}Y_k \sim \Gamma(n_m-1,1)$, and $1/\sum_{k=2}^{n_m}Y_k$ is inverse-gamma distribution with shape parameter $n_m - 1$ and rate parameter 1, we have $\mathbf{E}\frac{1}{\sum_{k=2}^{n_m}Y_k} = \frac{1}{n_m-2}$ for $n_m > 2$. Hence,

$$P(\sum_{k=2}^{n_m}\frac{1}{Y_k} > \epsilon) \leqslant \frac{(n_m-1)^2}{(n_m-2)\epsilon}.$$

Therefore by Lemma 7.8,

$$\sum_{k=2}^{n_m} \frac{1}{Y_k} = O_p\left(\frac{(n_m-1)^2}{n_m-2}\right) = O_p(n_m).$$

For (3.2),

$$\sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1-\rho^{2(t_k-t_{k-1})}} \mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]$$
$$= \sum_{k=2}^{n_m} \frac{\rho^{2(t_k-t_{k-1})}}{1+\rho^{t_k-t_{k-1}}} \frac{t_k-t_{k-1}}{1-\rho^{t_k-t_{k-1}}} \frac{\mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]}{t_k-t_{k-1}}. \quad (*)$$

It is easy to verify that there is constant L such that

$$\left|\frac{\rho^{2(t_k-t_{k-1})}}{1+\rho^{t_k-t_{k-1}}} - \frac{1}{2}\right| \leqslant L(t_k-t_{k-1}),$$

$$\left|\frac{t_k-t_{k-1}}{1-\rho^{t_k-t_{k-1}}} - \frac{1}{-\log\rho}\right| \leqslant L(t_k-t_{k-1}).$$

Considering

$$\frac{\mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]}{t_k-t_{k-1}}.$$

By the definition of C_{xz},

$$\frac{\mathbf{E}[(x(t_k)-x(t_{k-1}))(z(t_k)-z(t_{k-1}))|t_{k-1},t_k]}{t_k-t_{k-1}}$$
$$= \frac{\mathbf{E}[x(t_k)z(t_k)|t_{k-1},t_k] - \mathbf{E}[x(t_{k-1})z(t_k)|t_{k-1},t_k] - \mathbf{E}[x(t_k)z(t_{k-1})|t_{k-1},t_k] + \mathbf{E}[x(t_{k-1})z(t_{k-1})|t_{k-1},t_k]}{t_k-t_{k-1}}$$
$$= \frac{C_{xz}(t_k,t_k) - C_{xz}(t_{k-1},t_k) - C_{xz}(t_k,t_{k-1}) + C_{xz}(t_{k-1},t_{k-1})}{t_k-t_{k-1}}.$$

By Lemma 7.9, there is constant L such that

$$\left|\frac{C_{xz}(t_k,t_k) - C_{xz}(t_{k-1},t_k) - C_{xz}(t_k,t_{k-1}) + C_{xz}(t_{k-1},t_{k-1})}{t_k-t_{k-1}} - \psi_{x,z}(t_k)\right| \leqslant L(t_k - t_{k-1}),$$

and $\psi_{x,z}(t)$ is bounded on $[0, T]$. Therefore,

$$\begin{aligned}(*) &= \sum_{k=2}^{n_m} \left(\frac{1}{2} + O_p(t_k - t_{k-1})\right) \left(\frac{1}{-\log\rho} + O_p(t_k - t_{k-1})\right) (\psi_{x,z}(t_k) + O_p(t_k - t_{k-1})) \\ &= \sum_{k=2}^{n_m} -\frac{1}{2\log\rho}\psi_{x,z}(t_k) + \sum_{k=2}^{n_m} O_p(t_k - t_{k-1}) \\ &= -\frac{1}{2\log\rho}\sum_{k=2}^{n_m}\psi_{x,z}(t_k) + O_p(1) = -\frac{1}{2\log\rho}\sum_{k=1}^{n_m}\psi_{x,z}(t_k) + O_p(1)\end{aligned}$$

Hence (3.2) $= -\frac{1}{2\log\rho}\sum_{k=1}^{n_m}\psi_{x,z}(t_k) + O_p(1)$. Altogether, (1) $= O_p(1)$; (2.1) $= o_p(\frac{n_m}{g_1(n_s)})$, as $n_s \to \infty$; (2.2) $= O_p(1)$; (3.1) $= o_p(\frac{n_m^{1.5}}{g_2(n_s)})$, as $n_s \to \infty$; (3.2) $= -\frac{1}{2\log\rho}\sum_{k=1}^{n_m}\psi_{x,z}(t_k) + O_p(1)$. Let $g(x) = [\min\{g_1(x), g_2(x)\}]^{2/3}$, then

$$(2.1) + (3.1) = o_p(\frac{n_m^{1.5}}{g^{1.5}(n_s)}), \text{ as } n_s \to \infty.$$

Therefore as $n_s \to \infty$ and $\frac{n_m}{g(n_s)}$ is bounded,

$$\frac{1}{n_s}\sum_{j=1}^{n_s} X_j^T R^{-1} Z_j = -\frac{1}{2\sigma^2\log\rho}\sum_{k=1}^{n_m}\psi_{x,z}(t_k) + O_p(1).$$

□

Asymptotic property of $\tilde{X}^T Z$

Lemma 7.11. *For any two sequences,* $a_1, \ldots, a_n$ *and* $b_1, \ldots, b_n$*, we have* $\sum_{i=1}^n a_i b_i = M\sum_{i=1}^n |b_i|$*, where* $|M| \leqslant \max_{1\leqslant i\leqslant n}|a_i|$.

Proof. Because $M = \sum_{i=1}^n a_i b_i / \sum_{i=1}^n |b_i| = \sum_{i=1}^n a_i \operatorname{sign}(b_i)|b_i| / \sum_{i=1}^n |b_i|$ is a weighted average of $a_i \operatorname{sign}(b_i)$, $1 \leqslant i \leqslant n$, we have $\min_{1\leqslant i\leqslant n}\{a_i \operatorname{sign}(b_i)\} \leqslant M \leqslant \max_{1\leqslant i\leqslant n}\{a_i \operatorname{sign}(b_i)\}$, which implies $|M| \leqslant \max_{1\leqslant i\leqslant n}|a_i|$. □

Lemma 7.12. *Let* s *be a random variable with density* $f(s)$ *and* $g(t)$ *be a function defined on* $[0,T]$*. Both* f *and* g *have bounded derivatives. Then* $\forall 0<\delta<T/2$ *and for bounded* h*,*

$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{\mathbf{E}[(s-t)^r K_h(s-t)g(s)]-h^r\mu_r g(t)f(t)}{h^r}\right| = O(h),$$
$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{\mathrm{Var}[(s-t)^r K_h(s-t)g(s)]-h^{2r-1}\nu_r g^2(t)f(t)}{h^{2r-1}}\right| = O(h).$$

Proof. For $h<\delta$, $\forall t\in[\delta, T-\delta]$,

$$\begin{aligned}
\mathbf{E}(s-t)^r K_h(s-t)g(s) &= \int_0^T (s-t)^r K_h(s-t)g(s)f(s)\,ds\\
&= \int_{t-h}^{t+h} (s-t)^r K_h(s-t)g(s)f(s)\,ds\\
&= \int_{-1}^{1} (\nu h)^r \frac{K(\nu)}{h} g(t+\nu h)f(t+\nu h)\,d\nu h\\
&= h^r\int_{-1}^{1}\nu^r K(\nu)g(t)f(t)\,d\nu + h^r\int_{-1}^{1}\nu^r K(\nu)(gf)'(t+\tilde{\nu}h)\nu h\,d\nu\\
&= h^r\mu_r g(t)f(t) + h^{r+1}\int_{-1}^{1}\nu^r K(\nu)(gf)'(t+\tilde{\nu}h)\nu\,d\nu.
\end{aligned}$$

Because $(gf)'$ is bounded,

$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|\int_{-1}^{1}\nu^r K(\nu)(gf)'(t+\tilde{\nu}h)\nu\,d\nu\right| = O(1),$$

that is,

$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{\mathbf{E}[(s-t)^r K_h(s-t)g(s)]-h^r\mu_r g(t)f(t)}{h^{r+1}}\right| = O(1).$$

Now we consider the variance.

(i) From the above results, we get

$$\sup_{\delta\leqslant t\leqslant T-\delta}\frac{[\mathbf{E}(s-t)^r K_h(s-t)g(s)]^2}{h^{2r}}$$

$$
\begin{aligned}
&= \left[\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathbf{E}(s-t)^r K_h(s-t) g(s)}{h^r} \right| \right]^2 \\
&= \left[\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathbf{E}(s-t)^r K_h(s-t) g(s) - h^r \mu_r g(t) f(t)}{h^r} + \mu_r g(t) f(t) \right| \right]^2 \\
&\leqslant \left[\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathbf{E}(s-t)^r K_h(s-t) g(s) - h^r \mu_r g(t) f(t)}{h^r} \right| + \sup_{\delta \leqslant t \leqslant T-\delta} |\mu_r g(t) f(t)| \right]^2 \\
&= \left[O(h) + \sup_{\delta \leqslant t \leqslant T-\delta} |\mu_r g(t) f(t)| \right]^2 \\
&= [O(h) + O(1)]^2 = O(1) \quad \text{(since } h \text{ is bounded, we have } O(h) = O(1)\text{).} \qquad (*)
\end{aligned}
$$

(ii)

$$
\begin{aligned}
&\mathbf{E}(s-t)^{2r} K_h^2(s-t) g^2(s) \\
&= \int_0^T (s-t)^{2r} K_h^2(s-t) g^2(s) f(s)\, ds \\
&= \int_{-1}^1 (\nu h)^{2r} \frac{K^2(\nu)}{h^2} g^2(t) f(t)\, d\nu h + \int_{-1}^1 (\nu h)^{2r} \frac{K^2(\nu)}{h^2} (g^2 f)'(t + \tilde{\nu} h) \nu h\, d\nu h \\
&= h^{2r-1} \nu_r g^2(t) f(t) + h^{2r} \int_{-1}^1 \nu^{2r} K^2(\nu) (g^2 f)'(t + \tilde{\nu} h) \nu\, d\nu.
\end{aligned}
$$

Similarly,

$$
\sup_{\delta \leqslant t \leqslant T-\delta} \left| \int_{-1}^1 \nu^{2r} K^2(\nu) (g^2 f)'(t + \tilde{\nu} h) \nu\, d\nu \right| = O(1). \qquad (**)
$$

Using the fact

$$
\mathrm{Var}[(s-t)^r K_h(s-t) g(s)] = \mathbf{E}(s-t)^{2r} K_h^2(s-t) g^2(s) - [\mathbf{E}(s-t)^r K_h(s-t) g(s)]^2,
$$

and combining $(*)$ and $(**)$, we have

$$
\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathrm{Var}[(s-t)^r K_h(s-t)] - h^{2r-1} \nu_r g^2(t) f(t)}{h^{2r}} \right|
$$

$$
\begin{aligned}
&= \sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathbb{E}(s-t)^{2r} K_h^2(s-t) g^2(s) - [\mathbb{E}(s-t)^r K_h(s-t) g(s)]^2 - h^{2r-1} \nu_r g^2(t) f(t)}{h^{2r}} \right| \\
&= \sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{\mathbb{E}(s-t)^{2r} K_h^2(s-t) g^2(s) - h^{2r-1} \nu_r g^2(t) f(t)}{h^{2r}} - \frac{[\mathbb{E}(s-t)^r K_h(s-t) g(s)]^2}{h^{2r}} \right| \\
&= \sup_{\delta \leqslant t \leqslant T-\delta} \left| \int_{-1}^{1} \nu^{2r} K^2(\nu) (g^2 f)'(t + \tilde{\nu} h) \nu \, d\nu - \frac{[\mathbb{E}(s-t)^r K_h(s-t) g(s)]^2}{h^{2r}} \right| \\
&\leqslant \sup_{\delta \leqslant t \leqslant T-\delta} \left| \int_{-1}^{1} \nu^{2r} K^2(\nu) (g^2 f)'(t + \tilde{\nu} h) \nu \, d\nu \right| + \sup_{\delta \leqslant t \leqslant T-\delta} \frac{[\mathbb{E}(s-t)^r K_h(s-t) g(s)]^2}{h^{2r}} \\
&= O(1) + O(1) = O(1).
\end{aligned}
$$

□

Lemma 7.13. *Suppose random variables* $t_1, \ldots, t_n$ *are i.i.d. with density function* $f(t)$ *and distribution function* $F(t)$, *and* f *has a bounded derivative. Let* $\hat{F}_n(t)$ *be the empirical process, that is,* $\hat{F}_n(t) = \sum_{i=1}^n I_{t_i \leqslant t}$. *Then for* $s_1 < s_2$, *we have*

$$
\sup_{s \in [s_1, s_2]} \left| \frac{\hat{F}_n(s_2) - \hat{F}_n(s_1)}{s_2 - s_1} - f(s) \right| = O_p\left(\frac{1}{\sqrt{n}(s_2 - s_1)}\right) + O(s_2 - s_1).
$$

Proof. Let $s_1 \leqslant s \leqslant s_2$,

$$
\begin{aligned}
&\left| \frac{\hat{F}_n(s_2) - \hat{F}_n(s_1)}{s_2 - s_1} - f(s) \right| \\
=& \left| \frac{\hat{F}_n(s_2) - F(s_2)}{s_2 - s_1} - \frac{\hat{F}_n(s_1) - F(s_1)}{s_2 - s_1} + \frac{F(s_2) - F(s_1)}{s_2 - s_1} - f(s) \right| \\
=& \left| \frac{\hat{F}_n(s_2) - F(s_2)}{s_2 - s_1} - \frac{\hat{F}_n(s_1) - F(s_1)}{s_2 - s_1} + f(\tilde{s}) - f(s) \right| \\
\leqslant& 2 \sup_{0 \leqslant s \leqslant T} \left| \frac{\hat{F}_n(s) - F(s)}{s_2 - s_1} \right| + \sup_{|s - \tilde{s}| \leqslant s_2 - s_1} |f(\tilde{s}) - f(s)|.
\end{aligned}
$$

(i)

$$
\sup_{0 \leqslant s \leqslant T} \left| \frac{\hat{F}_n(s) - F(s)}{s_2 - s_1} \right| = \frac{1}{\sqrt{n}(s_2 - s_1)} \sup_{0 \leqslant s \leqslant T} \left| \sqrt{n}(\hat{F}_n(s) - F(s)) \right|.
$$

According to Kolmogorov's theorem, $\sup_{0\leqslant s\leqslant T}|\sqrt{n}(\hat{F}_n(s)-F(s))|$ converges in distribution to $\sup_{0\leqslant t\leqslant T} B(F(s))$, where $B(s)$ is a Brownian bridge on the unit interval. Therefore,

$$\sup_{0\leqslant s\leqslant T}\left|\frac{\hat{F}_n(s)-F(s)}{s_2-s_1}\right| = \frac{1}{\sqrt{n}(s_2-s_1)}O_p(1).$$

(ii) Since f' is bounded,

$$\sup_{|s-\tilde{s}|\leqslant s_2-s_1}|f(\tilde{s})-f(s)| \leqslant (s_2-s_1)\sup_{0\leqslant t\leqslant T}|f'(s)|.$$

Combining (i) and (ii), we complete the proof of the Lemma. □

Lemma 7.14. *Let* t_i, $i=1,\dots,n$ *be i.i.d with density* $f(t)$ *on* $[0,T]$ *and* g *be a function on* $[0,T]$. *Both* f *and* g *have bounded derivatives. Let*

$$B_{n,h}(t)=\frac{1}{nh}\sum_{k=1}^{n}\left((\frac{t_k-t}{h})^r K(\frac{t_k-t}{h})g(t_k)-E\left[(\frac{t_1-t}{h})^r K(\frac{t_1-t}{h})g(t_1)\right]\right).$$

Then $\forall 0<\delta<T/2$, *as* $n\to\infty$ *and* $h\to 0+$,

$$\sup_{\delta\leqslant t\leqslant T-\delta}|B_{n,h}(t)| = o_p(\frac{1}{n^{1/4}h^{1/2}})+o_p(\frac{1}{nh}).$$

In particular, if further we impose $1/h=O(\sqrt{n})$, *then* $\sup_{\delta\leqslant t\leqslant T-\delta}|B_{n,h}(t)|=o_p(1)$.

Proof. Let L be an integer. Let $s_1<\cdots<s_L$ be equally spaced on $[\delta, T-\delta]$.

(1) We first consider $\max_{1\leqslant l\leqslant L}|B_{n,h}(s_l)|$. Fix $\epsilon>0$,

$$P(\max_{1\leqslant l\leqslant L}|B_{n,h}(s_l)|>\epsilon)\leqslant\sum_{l=1}^{L}P(|B_{n,h}(s_l)|>\epsilon)\leqslant\sum_{l=1}^{L}\frac{\mathbf{E}[B_{n,h}^2(s_l)]}{\epsilon^2}.$$

According to Lemma 7.12, as $h\to 0+$,

$$\begin{aligned}\mathbf{E}[B_{n,h}^2(s_l)] =& \frac{1}{nh^{2r}}\operatorname{Var}[(t_1-s_l)^r K_h(t_1-s_l)g(t_1)]\\ =&\frac{1}{nh^{2r}}h^{2r-1}(\nu_r f(s_l)g^2(s_l)+o(1))\end{aligned}$$

$$=\frac{1}{nh}(\nu_r f(s_l)+o(1)),$$

Hence, we have

$$\begin{aligned}\sum_{l=1}^{L}\frac{E[B_{n,h}^2(s_l)]}{\epsilon^2}=&\sum_{l=1}^{L}\frac{\nu_r f(s_l)g^2(s_l)+o(1)}{nh\epsilon^2}\\ =&\frac{L\nu_r}{nh\epsilon^2}\frac{\sum_{l=1}^{L}(f(s_l)g^2(s_l)+o(1))}{L}\\ =&\frac{L\nu_r}{nh\epsilon^2}(\sup_s f(s)g^2(s)+o(1)).\end{aligned}$$

Therefore, $\max_{1\leqslant l\leqslant L}|B_{n,h}(s_l)|=O_p(\sqrt{\frac{L}{nh}})$.

(2) Now we prove that for $s_l\leqslant s\leqslant s_{l+1}$, $B_{n.h}(s)$ is not quite different from $B_{n.h}(s_l)$ or $B_{n.h}(s_{l+1})$.

$$\begin{aligned}&|B_{n.h}(s)-B_{n.h}(s_l)|\\ \leqslant&\frac{1}{nh}\sum_{k=1}^{n}\Bigg|[(\frac{t_k-s_l}{h})^rK(\frac{t_k-s_l}{h})g(t_k)-E(\frac{t_1-t}{h})^rK(\frac{t_1-t}{h})g(t_1)]\\ &-[(\frac{t_k-s}{h})^rK(\frac{t_k-s}{h})g(t_k)-E(\frac{t_1-t}{h})^rK(\frac{t_1-t}{h})g(t_1)]\Bigg|\\ =&\frac{1}{nh}\sum_{k=1}^{n}\Bigg|(\frac{t_k-s_l}{h})^rK(\frac{t_k-s_l}{h})g(t_k)-(\frac{t_k-s}{h})^rK(\frac{t_k-s}{h})g(t_k)\Bigg|\\ \leqslant&\frac{1}{nh}\sum_{k=1}^{n}(\sup_u[u^rK(u)]')|g(t_k)|\left|\frac{t_k-s_l}{h}-\frac{t_k-s}{h}\right|I_{(s-h<t_k<s_l+h)}\\ &+\frac{1}{nh}\sum_{k=1}^{n}(\sup_{-1\leqslant u\leqslant 1}u^rK(u))\sup_{0\leqslant u\leqslant T}g(u)(I_{(s_l-h<t_k\leqslant s-h)}+I_{(s_l+h\leqslant t_k<s+h)}).\end{aligned}$$

Since

$$\left|\frac{t_k-s_l}{h}-\frac{t_k-s}{h}\right|=\left|\frac{s-s_l}{h}\right|\leqslant\frac{T-2h}{h(L-1)}\sim O(\frac{1}{hL}),$$

we have

$$|B_{n.h}(s) - B_{n.h}(s_l)| \leqslant \frac{1}{nh} O(\frac{1}{hL}) \sum_{k=1}^{n} I_{(s-h<t_k<s_l+h)} + \frac{1}{nh} (\sup_{-1\leqslant u\leqslant 1} u^r K(u)) \sup_{0\leqslant u\leqslant T} g(u) \sum_{k=1}^{n} (I_{(s_l-h<t_k\leqslant s-h)} + I_{(s_l+h\leqslant t_k<s+h)}).$$

According to Lemma 7.13,

$$\sum_{k=1}^{n} I_{(s-h<t_k<s_l+h)} \leqslant \sum_{k=1}^{n} I_{(s_l-h<t_k<s_l+h)} = 2h(f(s_l) + O_p(\frac{1}{h\sqrt{n}}) + O(h)),$$
$$\sum_{k=1}^{n} I_{(s_l-h<t_k\leqslant s-h)} \leqslant \sum_{k=1}^{n} I_{(s_l-h<t_k\leqslant s_{l+1}-h)} = O(\frac{1}{L})(f(s_l) + O_p(\frac{L}{\sqrt{n}}) + O(\frac{1}{L})),$$
$$\sum_{k=1}^{n} I_{(s_l+h<t_k\leqslant s+h)} \leqslant \sum_{k=1}^{n} I_{(s_l+h<t_k\leqslant s_{l+1}+h)} = O(\frac{1}{L})(f(s_l) + O_p(\frac{L}{\sqrt{n}}) + O(\frac{1}{L})).$$

Hence as $L \to \infty$ and $L/\sqrt{n} \to 0$, we have

$$\max_{1\leqslant l\leqslant L-1} \sup_{s_l\leqslant s\leqslant s_{l+1}} |B_{n.h}(s) - B_{n.h}(s_l)| = O_p(\frac{1}{nhL}).$$

Combining (1) and (2), as $n \to \infty$, we have

$$\begin{aligned}
\sup_{\delta\leqslant t\leqslant T-\delta} |B_{n,h}(t)| \leqslant& \max_{1\leqslant l\leqslant L} |B_{n,h}(s_l)| + \max_{1\leqslant l\leqslant L-1} \sup_{s_l\leqslant s\leqslant s_{l+1}} |B_{n.h}(s) - B_{n.h}(s_l)| \\
=& O_p(\sqrt{\frac{L}{nh}}) + O_p(\frac{1}{nhL}) \\
=& \frac{1}{n^{1/4}h^{1/2}} O_p((L/\sqrt{n})^{1/2}) + \frac{1}{nh} O_p(\frac{1}{L}) \\
=& \frac{1}{n^{1/4}h^{1/2}} o_p(1) + \frac{1}{nh} o_p(1) \quad \text{(because } L/\sqrt{n} \to 0 \text{ and } L \to \infty) \\
=& o_p(\frac{1}{n^{1/4}h^{1/2}}) + o_p(\frac{1}{nh}).
\end{aligned}$$

If further we impose $1/h = O(\sqrt{n})$, or $\frac{1}{h\sqrt{n}} = O(1)$, then

$$o_p(\frac{1}{n^{1/4}h^{1/2}}) + o_p(\frac{1}{nh}) = (\frac{1}{h\sqrt{n}})^{1/2} o_p(1) + o_p(\frac{1}{nh})$$

$$
\begin{aligned}
&= O_p(1)o_p(1) + o_p(\frac{1}{nh}) \\
&= o_p(1) + o_p(\frac{1}{nh}) \\
&= o_p(1) + \frac{1}{h\sqrt{n}} o_p(\frac{1}{\sqrt{n}}) \\
&= o_p(1) + O_p(1)o_p(1) = o_p(1).
\end{aligned}
$$

□

It is known that (Wand and Jones 1994, pp.123)

$$
n_m^{-1} \sum_{k=1}^{n_m} (t_k - t)^r K_h(t_k - t) = \begin{cases} h^r \mu_r f(t) + o_P(h^r) & l \text{ even,} \\ h^{r+1} \mu_{r+1} f'(t) + o_P(h^{r+1}) & l \text{ odd.} \end{cases}
$$

We generalize this result in two directions. First, we consider the asymptotic property of the sum of $(t_k - t)^r K_h(t_k - t) x_{ijk} x_{ijk}^T$. Second, we consider the supremum when t varies in an interval, rather than a fixed t.

Lemma 7.15. *Recall that we defined*

$$
s_r(t, h) = \frac{1}{n_c n_s n_m} \sum_{i=1, j=1, k=1}^{n_c, n_s, n_m} (t_k - t)^r K_h(t_k - t) x_{ijk} x_{ijk}^T,
$$

where $t_1, \ldots, t_{n_m}$ *are i.i.d. with density function* f, $x_{ijk} = x_{ij}(t_k)$, *and* x_{ij}*s are replicates of a stochastic process* x. *If* f *has a bounded derivative and* $Ex(t)x(t)^T$ *as a function of* t *has a bounded derivative, then* $\forall\, 0 < \delta < T/2$, *as* $n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, *and* $n_c n_s \to \infty$, *elementwisely we have*

$$
\sup_{\delta \leqslant t \leqslant T - \delta} \left| \frac{s_r(t, h) - h^r \mu_r f(t) Ex(t)x(t)^T}{h^r} \right| = o_p(1).
$$

Note: Using a similar proof, this lemma could be generalized to the following statement:

Suppose $t_1, \dots, t_{n_m}$ *are i.i.d. with density function* f. *Let* $x_{jk} = x_j(t_k)$, *where* x_j*'s are replicates of a stochastic process* x. *Let* $z_{jk} = z_j(t_k)$, *where* z_j*'s are replicates of a stochastic process* z. *If* f *has a bounded derivative and* $\mathbf{E}x(t)z(t)^T$ *as a function of* t *has a bounded derivative, then* $\forall\, 0 < \delta < T/2$, *as* $n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, *and* $n_s \to \infty$,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{h^r n_s n_m} \sum_{j=1,k=1}^{n_s,n_m} (t_k - t)^r K_h(t_k - t) x_{jk} z_{jk}^T - \mu_r f(t) \mathbf{E}x(t)z(t)^T \right| = o_p(1).$$

Proof.

$$\begin{aligned} s_r(t,h) =& \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} (t_k - t)^r K_h(t_k - t) x_{ijk} x_{ijk}^T \\ =& \frac{1}{n_m} \sum_{k=1}^{n_m} \left\{ (t_k - t)^r K_h(t_k - t) \frac{1}{n_c n_s} \sum_{i=1,j=1}^{n_c,n_s} x_{ijk} x_{ijk}^T \right\} \\ =& \frac{1}{n_m} \sum_{k=1}^{n_m} (t_k - t)^r K_h(t_k - t) \mathbf{E}[x(t_k)x(t_k)^T | t_k] \qquad (1) \\ &+ \frac{1}{n_m} \sum_{k=1}^{n_m} (t_k - t)^r K_h(t_k - t) \left(\frac{1}{n_c n_s} \sum_{i=1,j=1}^{n_c,n_s} x_{ijk} x_{ijk}^T - \mathbf{E}[x(t_k)x(t_k)^T | t_k] \right). \qquad (2) \end{aligned}$$

It can be shown that $\sup_{\delta \leqslant t \leqslant T-\delta} |(1) - h^r \mu_r f(t) \mathbf{E}x(t)x(t)^T| = o_p(h^r)$, and $\sup_{\delta \leqslant t \leqslant T-\delta} |(2)| = o_p(h^r)$.

(i) Since

$$\begin{aligned} &\frac{1}{n_m} \sum_{k=1}^{n_m} (t_k - t)^r K_h(t_k - t) \mathbf{E}[x(t_k)x(t_k)^T | t_k] \\ =& \mathbf{E}((t_1 - t)^r K_h(t_1 - t) \mathbf{E}[x(t_1)x(t_1)^T | t_1]) \\ &+ \frac{1}{n_m} \sum_{k=1}^{n_m} ((t_k - t)^r K_h(t_k - t) - \mathbf{E}((t_1 - t)^r K_h(t_1 - t) \mathbf{E}[x(t_1)x(t_1)^T | t_1]), \end{aligned}$$

we have

$$
\begin{aligned}
&\sup_{\delta\leqslant t\leqslant T-\delta} |(1) - \mu_r f(t)\mathbf{E}x(t)x(t)^\mathsf{T}| \\
&= \sup_{\delta\leqslant t\leqslant T-\delta} \Big|(1) - \mathbf{E}((t_1-t)^r K_h(t_1-t)\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}|t_1]) + \\
&\quad + \mathbf{E}((t_1-t)^r K_h(t_1-t)\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}|t_1]) - \mu_r f(t)\mathbf{E}x(t)x(t)^\mathsf{T}\Big| \\
&\leqslant \sup_{\delta\leqslant t\leqslant T-\delta} \Big|(1) - \mathbf{E}((t_1-t)^r K_h(t_1-t)\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}|t_1])\Big| + \qquad (1.1) \\
&\quad + \sup_{\delta\leqslant t\leqslant T-\delta} \Big|\mathbf{E}((t_1-t)^r K_h(t_1-t)\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}|t_1]) - h^r\mu_r f(t)\mathbf{E}x(t)x(t)^\mathsf{T}\Big|. \qquad (1.2)
\end{aligned}
$$

By Lemma 7.12,

$$(1.2) = \sup_{\delta\leqslant t\leqslant T-\delta} \Big|\mathbf{E}((t_1-t)^r K_h(t_1-t)\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}]) - h^r\mu_r f(t)\mathbf{E}[x(t)x(t)^\mathsf{T}]\Big| = h^r O(h).$$

By Lemma 7.14, as $n_m \to \infty$ and $1/h = O(\sqrt{n_m})$, elementwisely

$$\sup_{\delta\leqslant t\leqslant T-\delta} \left|\frac{1}{n_m}\sum_{k=1}^{n_m}(t_k-t)^r K_h(t_k-t)\mathbf{E}[x(t_k)x(t_k)^\mathsf{T}|t_k] - \mathbf{E}(t_1-t)^r K_h(t_1-t))\mathbf{E}[x(t_1)x(t_1)^\mathsf{T}]\right| = h^r o_p(1).$$

Hence as $n_m \to \infty$, $h \to 0+$, and $1/h = O(\sqrt{n_m})$, elementwisely

$$\sup_{\delta\leqslant t\leqslant T-\delta} \Big|(1) - h^r\mu_r f(t)\mathbf{E}[x(t)x(t)^\mathsf{T}]\Big| = h^r o_p(1).$$

(ii) By Lemma 7.11,

$$
\begin{aligned}
(2) =& \frac{1}{n_m}\sum_{k=1}^{n_m}(t_k-t)^r K_h(t_k-t)\left(\frac{1}{n_c n_s}\sum_{i=1,j=1}^{n_c,n_s} x_{ijk}x_{ijk}^\mathsf{T} - \mathbf{E}[x(t_k)x(t_k)^\mathsf{T}|t_k]\right) \\
=& M\frac{1}{n_m}\sum_{k=1}^{n_m}|t_k-t|^r K_h(t_k-t),
\end{aligned}
$$

where M is a $p \times p$ matrix satisfying the following condition elementwisely

$$|M| \leqslant \max_{1 \leqslant k \leqslant n_m} \left| \frac{1}{n_c n_s} \sum_{i=1,j=1}^{n_c,n_s} x_{ijk} x_{ijk}^T - \mathbf{E}[x(t_k)x(t_k)^T | t_k] \right|$$
$$\leqslant \sup_{0 \leqslant t \leqslant T} \left| \frac{1}{n_c n_s} \sum_{i=1,j=1}^{n_c,n_s} x_{ij}(t) x_{ij}^T(t) - \mathbf{E}x(t)x(t)^T \right|.$$

By Lemma 7.3, as $n_c n_s \to \infty$, $M = o_p(1)$. Similar to the proof of (i), $\frac{1}{n_m} \sum_{k=1}^{n_m} |t_k - t|^r K_h(t_k - t) = O_p(h^r)$, therefore,

$$\frac{1}{n_m} \sum_{k=1}^{n_m} (t_k - t)^r K_h(t_k - t) \left(\frac{1}{n_c n_s} \sum_{i=1,j=1}^{n_c,n_s} x_{ijk} x_{ijk}^T - \mathbf{E}[x(t_k)x(t_k)^T | t_k] \right) = o_p(h^r).$$

Combining (i) and (ii), we prove the lemma. □

Lemma 7.16. *Under the conditions of Lemma 7.15, for each* t,

$$\frac{s_r(t,h)}{h^r} - \mathbf{E}\frac{s_r(t,h)}{h^r} = O_p(\frac{1}{\sqrt{n_m h}}).$$

Proof. The proof is similar to that of lemma 7.15 (omitted). The $O_p(\frac{1}{\sqrt{n_m h}})$ term comes from the central limit theory. □

The following is a lemma of uniform convergence of stochastic processes.

Lemma 7.17. *Suppose* $X_n = \{X_n(t) : t \in [0,T]\}, n = 1, 2, \ldots$ *are stochastic processes,* $a(t)$ *is a real-valued function defined on* $[0,T]$*, and as* $n \to \infty$, $\sup_t |X_n(t) - a(t)| = o_p(1)$. *If* $\inf_t a(t) > 0$*, then*

$$\sup_t \left| \frac{1}{X_n(t)} - \frac{1}{a(t)} \right| = o_p(1).$$

Proof. Let $c = \inf_t a(t)$. For $\epsilon > 0$,

$$P(\sup_t \left| \frac{1}{X_n(t)} - \frac{1}{a(t)} \right| > \epsilon)$$

$$\leqslant P(\sup_t \left|\frac{1}{X_n(t)} - \frac{1}{a(t)}\right| > \epsilon, \sup_t |X_n(t) - a(t)| \leqslant c/2) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2)$$
$$\leqslant P(\sup_t \left|\frac{X_n(t) - a(t)}{X_n(t)a(t)}\right| > \epsilon, \sup_t |X_n(t) - a(t)| \leqslant c/2) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2)$$
$$\leqslant P(\sup_t \left|\frac{X_n(t) - a(t)}{(a(t) - c/2)a(t)}\right| > \epsilon, \sup_t |X_n(t) - a(t)| \leqslant c/2) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2)$$
$$\leqslant P(\sup_t \left|\frac{X_n(t) - a(t)}{(a(t) - c/2)a(t)}\right| > \epsilon) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2)$$
$$\leqslant P(\sup_t \left|\frac{X_n(t) - a(t)}{(c - c/2)c}\right| > \epsilon) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2)$$
$$\leqslant P(\sup_t |X_n(t) - a(t)| > \epsilon c^2/2) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2).$$

Since $\sup_t |X_n(t) - a(t)| = o_p(1)$,

$$P(\sup_t |X_n(t) - a(t)| > \epsilon c^2/2) + P(\sup_t |X_n(t) - a(t)| \geqslant c/2) = 0,$$

we have $P(\sup_t \left|\frac{1}{X_n(t)} - \frac{1}{a(t)}\right| > \epsilon) = 0$, i.e., $\sup_t \left|\frac{1}{X_n(t)} - \frac{1}{a(t)}\right| = o_p(1)$. □

Corollary 7.18. *Suppose* $X_n = \{X_n(t) : t \in [0, T]\}, n = 1, 2, \ldots$ *are stochastic processes, and the dimension of* $X_n(t)$ *is* p *by* p. $a(t)$ *is a matrix-valued function defined on* $[0, T]$ *with dimension* p *by* p*, and as* $n \to \infty$*, elementwisely* $\sup_t |X_n(t) - a(t)| = o_p(1)$*. If* $\inf_t \det(a(t)) > 0$*, then elementwisely*

$$\sup_t \left|X_n^{-1}(t) - a^{-1}(t)\right| = o_p(1).$$

Proof. The detailed proof is omitted. The idea is to use Cramer's rule and Lemma 7.17. □

Corollary 7.19. *If* f *has a bounded derivative and* $Ex(t)x(t)^T$ *as a function of* t *has a bounded derivative,* $\inf_t f(t) > 0$*, and* $\inf_t \det(Ex(t)x(t)^T) > 0$*, then* $\forall\, 0 < \delta < T/2$*, as*

$n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, *and* $n_c n_s \to \infty$, *elementwisely we have,*

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| [s_0(t,h) - s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1} - \frac{1}{f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \right| = o_p(1).$$

Proof. By Lemma 7.15,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{s_r(t,h) - h^r \mu_r f(t) \mathbf{E}x(t)x(t)^T}{h^r} \right| = o_p(1),$$

Hence,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| s_0(t,h) - f(t)\mathbf{E}x(t)x(t)^T \right| = o_p(1),$$

$$\sup_{\delta \leqslant t \leqslant T-\delta} |s_1(t,h)/h| = o_p(1),$$

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| s_2(t,h)/h^2 - \mu_2 f(t)\mathbf{E}x(t)x(t)^T \right| = o_p(1).$$

By Corollary 7.18,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| h^2[s_2(t,h)]^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \right| = o_p(1),$$

Hence,

$$\begin{aligned}
&\sup_{\delta \leqslant t \leqslant T-\delta} \left| s_1(t,h)s_2^{-1}(t,h)s_1(t,h) \right| \\
&= \sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{s_1(t,h)}{h} \left(h^2[s_2(t,h)]^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \right) \frac{s_1(t,h)}{h} + \right. \\
&\quad \left. \frac{s_1(t,h)}{h} \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \frac{s_1(t,h)}{h} \right| \\
&\leqslant \sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{s_1(t,h)}{h} \left(h^2[s_2(t,h)]^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \right) \frac{s_1(t,h)}{h} \right| + \qquad (1) \\
&\quad + \sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{s_1(t,h)}{h} \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1} \frac{s_1(t,h)}{h} \right| \qquad (2)
\end{aligned}$$

For (1),

$$\begin{aligned}
&\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\left(h^2[s_2(t,h)]^{-1}-\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1}\right)\frac{s_1(t,h)}{h}\right|\\
\leqslant&\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\right|\sup_{\delta\leqslant t\leqslant T-\delta}\left|h^2[s_2(t,h)]^{-1}-\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1}\right|\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\right|\\
\leqslant& o_p(1)o_p(1)o_p(1)=o_p(1);
\end{aligned}$$

For (2), similarly

$$\begin{aligned}
&\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1}\frac{s_1(t,h)}{h}\right|\\
\leqslant&\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\right|\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1}\right|\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1(t,h)}{h}\right|\\
=& o_p(1)O_p(1)o_p(1)=o_p(1),
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\delta\leqslant t\leqslant T-\delta}\left|[s_0(t,h)-s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]-f(t)\mathbf{E}x(t)x(t)^\top\right|\\
\leqslant&\sup_{\delta\leqslant t\leqslant T-\delta}\left|s_0(t,h)-f(t)\mathbf{E}x(t)x(t)^\top\right|+\sup_{\delta\leqslant t\leqslant T-\delta}\left|s_1(t,h)s_2^{-1}(t,h)s_1(t,h)\right|\\
=& o_p(1)+o_p(1)=o_p(1).
\end{aligned}$$

Again by Corollary 7.18,

$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|[s_0(t,h)-s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1}-\frac{1}{f(t)}[\mathbf{E}x(t)x(t)^\top]^{-1}\right|=o_p(1).$$

□

Corollary 7.20. *Under the conditions of Lemma 7.15, given* t, *if* $\mathbf{E}x(t)x(t)^\top$ *is invertible,* *then as* $h \to 0_+$, $n_m h \to \infty$,

$$[s_0(t,h) - s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1} - \frac{1}{f(t)}[\mathbb{E}x(t)x(t)^T]^{-1} = O_p(h) + O_p(\frac{1}{\sqrt{n_m h}}).$$

Proof. We prove this corollary using lemma 7.16.

$$\begin{aligned}
& s_0 - s_1 s_2^{-1} s_1 \\
=& s_0 - \frac{s_1}{h}\left[\frac{s_2}{h^2}\right]^{-1}\frac{s_1}{h} \\
=& f(t)\mathbb{E}x(t)x(t)^T + O(h) + O_p(\frac{1}{\sqrt{n_m h}}) \\
& + \left(O(h) + O_p(\frac{1}{\sqrt{n_m h}})\right)\left[\mu_2 f(t)\mathbb{E}x(t)x(t)^T + O(h) + O_p(\frac{1}{\sqrt{n_m h}})\right]^{-1}\left(O(h) + O_p(\frac{1}{\sqrt{n_m h}})\right) \\
=& f(t)\mathbb{E}x(t)x(t)^T + O(h) + O_p(\frac{1}{\sqrt{n_m h}}) \\
& + \left(O(h) + O_p(\frac{1}{\sqrt{n_m h}})\right)\left(\frac{1}{\mu_2 f(t)}[\mathbb{E}x(t)x(t)^T]^{-1} + O_p(h) + O_p(\frac{1}{\sqrt{n_m h}})\right)\left(O(h) + O_p(\frac{1}{\sqrt{n_m h}})\right) \\
=& f(t)\mathbb{E}x(t)x(t)^T + O_p(h) + O_p(\frac{1}{\sqrt{n_m h}}).
\end{aligned}$$

Therefore,

$$[s_0 - s_1 s_2^{-1} s_1]^{-1} = \frac{1}{f(t)}[\mathbb{E}x(t)x(t)^T]^{-1} + O_p(h) + O_p(\frac{1}{\sqrt{n_m h}}).$$

□

Lemma 7.21. *Suppose* $f(t)$, $\mathbb{E}x(t)x(t)^T$ *and* $\mathbb{E}x(t)z(t)^T$ *as functions of* t *have bounded derivatives, and* $\inf_t f(t) > 0$, $\inf_t \det(\mathbb{E}x(t)x(t)^T) > 0$. *As* $n_s \to \infty$, $n_m \to \infty$, $h \to 0+$ *and* $1/h = O(\sqrt{n_m})$,

$$\sup_{\delta \leqslant t \leqslant T-\delta}\left|\frac{1}{n_s n_m}\tilde{X}_i^T Z_i - f(t)\mathbb{E}x(t)z(t)^T\right| = o_p(1).$$

Proof. Recall that $Z = \mathrm{diag}(Z_1, \ldots, Z_{n_c})$, where

$$Z_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{in_s} \end{pmatrix}, \quad Z_{ij} = \begin{pmatrix} z_{ij1}^T \\ \vdots \\ z_{ijn_m}^T \end{pmatrix}.$$

$\tilde{X}$ is an abbreviation of $\tilde{X}(t, h)$, hence

$$\begin{aligned} \tilde{X}_i^T Z_i &= \sum_{j,k} \tilde{x}_{ijk} z_{ijk}^T = \sum_{j,k} (I_p - s_1 s_2^{-1}(t_k - t)) K_h(t_k - t) x_{ijk} z_{ijk}^T \\ &= \sum_{j,k} K_h(t_k - t) x_{ijk} z_{ijk}^T - s_1 s_2^{-1} \sum_{j,k} (t_k - t) K_h(t_k - t) x_{ijk} z_{ijk}^T. \end{aligned}$$

By Lemme 7.15, $\forall 0 < \delta < T/2$, as $n_s \to \infty$, $n_m \to \infty$, $h \to 0+$, and $1/h = O(\sqrt{n_m})$,

$$\begin{aligned} &\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{n_s n_m} \sum_{j,k} K_h(t_k - t) x_{ijk} z_{ijk}^T - f(t) \mathbf{E} x(t) z(t)^T \right| = o_p(1), \\ &\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{h n_s n_m} \sum_{j,k} (t_k - t) K_h(t_k - t) x_{ijk} z_{ijk}^T \right| = o_p(1), \\ &\sup_{\delta \leqslant t \leqslant T-\delta} |s_1(t, h)/h| = o_p(1), \\ &\sup_{\delta \leqslant t \leqslant T-\delta} \left| s_2(t, h)/h^2 - \mu_2 f(t) \mathbf{E} x(t) x(t)^T \right| = o_p(1). \end{aligned}$$

By Lemma 7.17, as $h \to 0+$,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| h^2 [s_2(t, h)]^{-1} - \frac{1}{\mu_2 f(t)} [\mathbf{E} x(t) x(t)^T]^{-1} \right| = o_p(1),$$

Therefore,

$$\begin{aligned} &\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{n_s n_m} \tilde{X}_i^T Z_i - f(t) \mathbf{E} x(t) z(t)^T \right| \\ = &\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{1}{n_s n_m} \left(\sum_{j,k} K_h(t_k - t) x_{ijk} z_{ijk}^T - s_1 s_2^{-1} \sum_{j,k} (t_k - t) K_h(t_k - t) x_{ijk} z_{ijk}^T \right) - f(t) \mathbf{E} x(t) z(t)^T \right| \end{aligned}$$

$$= \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{n_s n_m}\left(\sum_{j,k} K_h(t_k-t)x_{ijk}z_{ijk}^T - f(t)\mathbf{E}x(t)z(t)^T\right) - \right.$$

$$\left.\frac{s_1}{h}(h^2 s_2^{-1})\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|$$

$$\leqslant \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{n_s n_m}\sum_{j,k} K_h(t_k-t)x_{ijk}z_{ijk}^T - f(t)\mathbf{E}x(t)z(t)^T\right| - \tag{1}$$

$$- \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}(h^2 s_2^{-1})\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|. \tag{2}$$

We have shonw that $(1) = o_p(1)$. Now consider (2).

$$\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}(h^2 s_2^{-1})\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|$$

$$\leqslant \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}(h^2 s_2^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1})\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right| +$$

$$+ \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|$$

$$\leqslant \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}\right| \sup_{\delta\leqslant t\leqslant T-\delta}\left|h^2 s_2^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}\right|$$

$$\times \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|$$

$$+ \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{s_1}{h}\right| \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}\right| \sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{hn_s n_m}\sum_{j,k}(t_k-t)K_h(t_k-t)x_{ijk}z_{ijk}^T\right|$$

$$\leqslant o_p(1)o_p(1)o_p(1) + o_p(1)O_p(1)o_p(1) = o_p(1).$$

To sum up, $\sup_{\delta\leqslant t\leqslant T-\delta}\left|\frac{1}{n_s n_m}\tilde{X}_i^T Z_i - f(t)\mathbf{E}x(t)z(t)^T\right| = o_p(1)$. □

Corollary 7.22. *Under the conditions of Lemma 7.21,*

$$\frac{1}{n_s n_m}\tilde{X}_i^T Z_i - E(K_h(t_1-t)E(x(t_1)z(t_1)^T|t_1)) = O_p(\frac{1}{\sqrt{n_m h}}).$$

Proof. The proof is similar to the proof of Lemma 7.21 and is ommited. □

Asymptotic property of $Z^T(I - S_d)^T R_d^{-1}(I - S_d)Z$

Consider the matrix, $A_n = I_n - \frac{a}{n}J_n$, where J_n is n by n matrix of ones. A simple calculation reveals that $\det(A_n) = 1-a$, and thus $\lim_{n\to\infty}\det(A_n) = 1-a$. Consider a general case,

$$A_{nm} = I_n - \frac{1}{n}\begin{pmatrix} a_{11}(m) & a_{12}(m) & \cdots & a_{1n}(m) \\ a_{21}(m) & a_{22}(m) & \cdots & a_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(m) & a_{n2}(m) & \cdots & a_{nn}(m) \end{pmatrix},$$

where as $m \to \infty$, $a_{ij}(m) = a + o_p(1)$. Can we claim that $\lim_{\substack{n\to\infty \\ m\to\infty}} \det(A_{nm}) = 1-a$? The answer depends on how fast $a_{ij}(m)$ converges in probability to a and the relationship between n and m. Intuitively, if n goes to infinity too fast relative to m goes to infinity, the claim should not be true. Later we will consider the case when $a = 0$.

Lemma 7.23. *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with zero mean and finite third moment. Let Z be a random variable with a finite mean. Define $X_n = n^{-1/2}(\sum_{i=1}^n Y_i + Z)$. Then there exist constants a, b, c and d, where $b > 0$, such that $\forall t > 0$*

$$P(|X_n| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n}} + \frac{d}{t\sqrt{n}}.$$

Proof. Denote $EY_1^2 = \sigma^2$,

$$\begin{aligned} P(|X_n| > t) =& P\left(\left|\frac{\sum_{i=1}^n Y_i}{\sqrt{n}} + \frac{Z}{\sqrt{n}}\right| > t\right) \\ \leqslant& P\left(\left|\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right| + \left|\frac{Z}{\sqrt{n}}\right| > t\right) \end{aligned}$$

$$\leqslant P\left(\left|\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}\right| > t/2 \text{ or } \left|\frac{Z}{\sqrt{n}}\right| > t/2\right)$$
$$\leqslant P\left(\left|\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}\right| > t/2\right) + P\left(\left|\frac{Z}{\sqrt{n}}\right| > t/2\right).$$

By the Berry-Esseen theorem (Berry, 1941; Esseen, 1942), there exists a constant c such that $\forall x$

$$\left|P\left(\frac{\sum_{i=1}^{n} Y_i}{\sigma\sqrt{n}} \leqslant x\right) - \Phi(x)\right| \leqslant \frac{c}{\sqrt{n}}.$$

Hence,

$$\begin{aligned}
P\left(\left|\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}\right| > t/2\right) =& P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} > t/2\right) + P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} < t/2\right) \\
=& 1 - \Phi(\frac{t/2}{\sigma}) + P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} > t/2\right) - (1 - \Phi(\frac{t/2}{\sigma})) \\
& + \Phi(-\frac{t/2}{\sigma}) + P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} < -t/2\right) - \Phi(-\frac{t/2}{\sigma}) \\
\leqslant& 1 - \Phi(\frac{t/2}{\sigma}) + \left|P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} > t/2\right) - (1 - \Phi(\frac{t/2}{\sigma}))\right| \\
& + \Phi(-\frac{t/2}{\sigma}) + \left|P\left(\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} < -t/2\right) - \Phi(-\frac{t/2}{\sigma})\right| \\
\leqslant& 1 - \Phi(\frac{t/2}{\sigma}) + \Phi(-\frac{t/2}{\sigma}) + \frac{2c}{\sqrt{n}}
\end{aligned}$$

Using the well-known Gaussian tail bound, there are constants a and b, where $b > 0$, such that

$$1 - \Phi(\frac{t/2}{\sigma}) + \Phi(-\frac{t/2}{\sigma}) \leqslant \exp(a - bt^2).$$

By the Markov's inequality,

$$P\left(\left|\frac{Z}{\sqrt{n}}\right| > t/2\right) \leqslant \frac{\mathbf{E}|Z|}{\sqrt{n}t/2}.$$

Therefore let $d = 2\mathbf{E}|Z|$, this completes the proof. □

Lemma 7.24. *Let*

$$D_{nm} = \begin{pmatrix} a_{11}(m) & a_{12}(m) & \cdots & a_{1n}(m) \\ a_{21}(m) & a_{22}(m) & \cdots & a_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(m) & a_{n2}(m) & \cdots & a_{nn}(m) \end{pmatrix}$$

be a random matrix, where $n \geqslant 1$ *and* $m \geqslant 1$. *If there are constants* a, b, c *and* d, *where* $b > 0$, *such that* $\forall t > 0$, $i \geqslant 1$, $j \geqslant 1$,

$$P(\sqrt{m}|a_{ij}(m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{m}} + \frac{d}{t\sqrt{m}},$$

then as $m \to \infty$ *and* $n^4 = o(m)$, *we have* $\det(I_n + \frac{1}{n}D_{nm}) = 1 + o_p(1)$. *Let* $\tilde{I}_n =$ *diag*$(1+a_1(m), \dots, 1+a_n(m))$, *where* $a_i(m)$ *are random variables. If further as* $m \to \infty$ *and* $n^4 = o(m)$, $\sum_{i=1}^{n} |a_i(m)| = o_p(1)$, *then*

$$\det(\tilde{I}_n + \frac{1}{n}D_{nm}) = 1 + o_p(1).$$

A sufficient condition of $\sum_{i=1}^{n} |a_i(m)| = o_p(1)$ *as* $m \to \infty$ *and* $n^4 = o(m)$ *is that*

$$P(\sqrt{m}|a_i(m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{m}} + \frac{d}{t\sqrt{m}}.$$

Proof. Let $t = m^{-1/4}$ and $n^4 = o(m)$. We first prove that as $m \to \infty$,

$$n^2 \sup_{i \geqslant 1, j \geqslant 1} P(|a_{ij}(m)| > \frac{t}{\sqrt{n}}) = o(1).$$

By assumptions,

$$\begin{aligned} & n^2 \sup_{i \geqslant 1, j \geqslant 1} P(|a_{ij}(m)| > \frac{t}{\sqrt{n}}) \\ =& n^2 \sup_{i \geqslant 1, j \geqslant 1} P(\sqrt{m}|a_{ij}(m)| > \frac{t\sqrt{m}}{\sqrt{n}}) \\ \leqslant& n^2 \left(\exp(a - b(\frac{t\sqrt{m}}{\sqrt{n}})^2) + \frac{c}{\sqrt{m}} + \frac{d}{(\frac{t\sqrt{m}}{\sqrt{n}})\sqrt{m}} \right) \end{aligned}$$

$$=n^2\left(\exp(a-\frac{bt^2m}{n})+\frac{c}{\sqrt{m}}+\frac{d\sqrt{n}}{tm}\right)$$
$$=n^2\left(\exp(a-\frac{b\sqrt{m}}{n})+\frac{c}{\sqrt{m}}+\frac{d\sqrt{n}}{m^{3/4}}\right)$$

Since $n^2=o(\sqrt{m})$, as $m\to\infty$ we have $\log n<n=o(\frac{\sqrt{m}}{n})$, so

$$n^2\exp(a-\frac{b\sqrt{m}}{n})=o(1). \tag{1}$$

Since $n^4=o(m)$, as $m\to\infty$ we have

$$n^2\frac{1}{\sqrt{m}}=\sqrt{\frac{n^4}{m}}=o(1), \tag{2}$$
$$n^2\frac{\sqrt{n}}{m^{3/4}}=\left(\frac{n^4}{m}\right)^{\frac{5}{8}}\frac{1}{m^{1/8}}=o(1), \tag{3}$$

(1), (2) and (3) imply that $n^2\sup_{i\geqslant 1,j\geqslant 1}P(|a_{ij}(m)|>\frac{t}{\sqrt{n}})=o(1)$. Since

$$P(\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|>\frac{t}{\sqrt{n}})\leqslant\sum_{i=1,j=1}^{i=n,j=n}P(|a_{ij}(m)|>\frac{t}{\sqrt{n}})$$
$$\leqslant n^2\sup_{i\geqslant 1,j\geqslant 1}P(a_{ij}(m)>\frac{t}{\sqrt{n}}),$$

as $m\to\infty$ and $n^4=o(m)$, we have, $P(\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|>\frac{t}{\sqrt{n}})\to 0$, in other words,

$$P(\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|\leqslant\frac{t}{\sqrt{n}})\to 1.$$

We claim that for $t\leqslant\sqrt{n}$, which is true since $t=m^{-1/4}$, $\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|\leqslant\frac{t}{\sqrt{n}}$ implies that

$$(1-\frac{t}{n})^n(1-\frac{t}{\sqrt{n}})\leqslant\det(I_n+\frac{1}{n}D_{nm})\leqslant\left(1+\frac{t+t^2}{n/2}\right)^{n/2}.$$

Since

$$\lim_{\substack{m\to\infty\\ n^4=o(m)}}(1-\frac{t}{n})^n(1-\frac{t}{\sqrt{n}})=1,\quad \lim_{\substack{m\to\infty\\ n^4=o(m)}}\left(1+\frac{t+t^2}{n/2}\right)^{n/2}=1,$$

for $\forall\epsilon>0$, we can find M, such that when $m>M$ and $n^4=o(m)$, we have $(1-\frac{t}{n})^n(1-\frac{t}{\sqrt{n}})\geqslant 1-\epsilon$ and $\left(1+\frac{t+t^2}{n/2}\right)^{n/2}\leqslant 1+\epsilon$, which implies that $1-\epsilon\leqslant \det(I_n+\frac{1}{n}D_{nm})\leqslant 1+\epsilon$. To summarize, if $\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|\leqslant\frac{t}{\sqrt{n}}$, then $\forall\epsilon>0$, we have as for sufficiently large m,

$$1-\epsilon\leqslant\det(I_n+\frac{1}{n}D_{nm})\leqslant 1+\epsilon.$$

Hence,

$$P\left(\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|\leqslant\frac{t}{\sqrt{n}}\right)\leqslant P\left(1-\epsilon\leqslant\det(I_n+\frac{1}{n}D_{nm})\leqslant 1+\epsilon\right).$$

Note that as $m\to\infty$ and $n^4=o(m)$, the left-hand side of the above inequality converges to 1, therefore the right-hand side also converges to 1, which is the result this lemma want to prove. Now we show that when $t\leqslant\sqrt{n}$, $\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|\leqslant\frac{t}{\sqrt{n}}$ implies that

$$(1-\frac{t}{n})^n(1-\frac{t}{\sqrt{n}})\leqslant\det(I_n+\frac{1}{n}D_{nm})\leqslant\left(1+\frac{t+t^2}{n/2}\right)^{n/2}.$$

We use the geometric interpretation of determination, i.e., the determinant is the oriented volume of the parallelepiped spanned by the column or row vectors. Let v_i be the ith row vector of $I_n+\frac{1}{n}D_{nm}$. With regard to the upper bound, by Hadamard's inequality,

$$\det(I_n+\frac{1}{n}D_{nm})\leqslant\prod_{i=1}^{n}\|v_i\|$$

$$
\begin{aligned}
&= \prod_{i=1}^{n}\left[(1+\frac{a_{ii}(m)}{n})^2 + \sum_{j\neq i}(\frac{a_{ij}(m)}{n})^2\right]^{1/2} \\
&\leqslant \prod_{i=1}^{n}\left[(1+\frac{\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|}{n})^2 + \sum_{j\neq i}(\frac{\max_{1\leqslant i,j\leqslant n}|a_{ij}(m)|}{n})^2\right]^{1/2} \\
&\leqslant \prod_{i=1}^{n}\left[(1+\frac{t/\sqrt{n}}{n})^2 + \sum_{j\neq i}(\frac{t/\sqrt{n}}{n})^2\right]^{1/2} \\
&= \left((1+\frac{t}{n^{1.5}})^2 + (n-1)\frac{t^2}{n^3}\right)^{n/2} \\
&\leqslant \left(1+\frac{2t}{n^{1.5}} + \frac{2t^2}{n^2}\right)^{n/2} \\
&\leqslant \left(1+\frac{t+t^2}{n/2}\right)^{n/2}.
\end{aligned}
$$

As for the lower bound, let u_i be the ith row vector of $\frac{1}{n}D_{nm}$, and e_i be a row vector with the ith element is 1 and the rest are 0, then $\det(I_n + \frac{1}{n}D_{nm})$ is the volume of the parallelepiped spanned by $e_i + u_i$, $i = 1, \ldots, n$. Let $e = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$ be the unit vector along the diagonal direction. We can show that the length of the projection of u_i onto e is no greater than t/n, i.e.,

$$
| < u_i, e > | = \left|\sum_{j=1}^{n}\frac{a_{ij}(m)}{n}\frac{1}{\sqrt{n}}\right| \leqslant \sum_{j=1}^{n}\frac{|a_{ij}(m)|}{n\sqrt{n}} \leqslant \sum_{j=1}^{n}\frac{t/\sqrt{n}}{n\sqrt{n}} = \frac{t}{n}
$$

and the length of the projection of u_i onto the plane perpendicular to e is no greater than t/n, i.e., by Pythagorean theorem,

$$
\begin{aligned}
\|u_i - < u_i, e > e\| &= \sqrt{\|u_i\|^2 - \| < u_i, e > e\|^2} \\
&\leqslant \|u_i\| = \sqrt{\sum_{j=1}^{n}[\frac{a_{ij}(m)}{n}]^2} \leqslant \sqrt{\sum_{j=1}^{n}[\frac{t/\sqrt{n}}{n}]^2} = \frac{t}{n}.
\end{aligned}
$$

Decompose e_i as $e_i = \alpha e + \beta e_i^{\perp}$, where

$$\begin{aligned} & e \perp e_i^{\perp}, \\ & \alpha =< e, e_i >= \frac{1}{\sqrt{n}}, \\ & \beta = \sqrt{1-\alpha^2} = \sqrt{1-\frac{1}{n}}. \end{aligned}$$

Denote $\gamma = \frac{t}{n}$. When $t \leqslant \sqrt{n}$, the volume of the parallelepiped spanned by $e_i + u_i$, $i = 1, \dots, n$, where $\| < u_i, e > e\| \leqslant \gamma$ and $\|u_i - < u_i, e > e\| \leqslant \gamma$, is no smaller than that when $u_i = -\gamma e - \gamma e_i^{\perp}$. That is, the determinant of D_n, of which the ith row is $e_i - \gamma e - \gamma e_i^{\perp}$, is a lower bound for $\det(I_n + \frac{1}{n} D_{nm})$. Since

$$e_i = \alpha e + \beta e_i^{\perp} \Rightarrow e_i^{\perp} = \frac{1}{\beta} e_i - \frac{\alpha}{\beta} e,$$

the ith row of the matrix D_n can be re-presented by

$$\begin{aligned} & e_i - \gamma e - \gamma e_i^{\perp} = e_i - \gamma e - \gamma(\frac{1}{\beta} e_i - \frac{\alpha}{\beta} e) \\ =& (1-\frac{\gamma}{\beta}) e_i - \gamma(1-\frac{\alpha}{\beta}) e, \end{aligned}$$

Hence D_n could be expressed as

$$D_n = (1-\frac{\gamma}{\beta}) I_n - \gamma(1-\frac{\alpha}{\beta}) \frac{1}{\sqrt{n}} J_n,$$

where J_n is a $n \times n$ matrix of all ones. Therefore,

$$\begin{aligned} \det(D_n) =& (1-\frac{\gamma}{\beta})^{n-1} \left((1-\frac{\gamma}{\beta}) - \gamma(1-\frac{\alpha}{\beta}) \sqrt{n} \right) \\ =& (1-\frac{\gamma}{\beta})^{n-1} \left((1-\frac{\gamma}{\beta}) - \gamma(1-\frac{1/\sqrt{n}}{\beta}) \sqrt{n} \right) \\ =& (1-\frac{\gamma}{\beta})^{n-1} (1-\sqrt{n}\gamma) = (1-\frac{t/n}{\sqrt{1-\frac{1}{n}}})^{n-1} (1-\sqrt{n}\frac{t}{n}) \end{aligned}$$

$$\geqslant (1-\frac{t}{n})^n(1-\frac{t}{\sqrt{n}}).$$

Now we consider $\det(\tilde{I}_n + \frac{1}{n}D_{nm})$. Since as $m \to \infty$ and $n^4 = o(m)$, $\sum_{i=1}^n |a_i| = o_p(1)$, we have $P(\min_{1\leqslant i\leqslant n} 1 + a_i > 1/2) \to 1$. Suppose $\min_{1\leqslant i\leqslant n} 1 + a_i > 1/2$, let

$$\tilde{a}_{ij}(m) = \frac{a_{ij}(m)}{\sqrt{1+a_i}\sqrt{1+a_j}}$$

and

$$\tilde{D}_{nm} = \begin{pmatrix} \tilde{a}_{11}(m) & \tilde{a}_{12}(m) & \cdots & \tilde{a}_{1n}(m) \\ \tilde{a}_{21}(m) & \tilde{a}_{22}(m) & \cdots & \tilde{a}_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{n1}(m) & \tilde{a}_{n2}(m) & \cdots & \tilde{a}_{nn}(m) \end{pmatrix},$$

then

$$\det(\tilde{I}_n + \frac{1}{n}D_{nm}) = \det(I_n + \frac{1}{n}\tilde{D}_{nm})\prod_{i=1}^n(1+a_i).$$

Since

$$\prod_{i=1}^n(1+a_i) = \exp(\sum_{i=1}^n \log(1+a_i)) \leqslant \exp(\sum_{i=1}^n a_i) = \exp(o(1)) = 1 + o(1),$$

we only need to prove that $\det(I_n + \frac{1}{n}\tilde{D}_{nm}) = 1 + o_p(1)$. Since

$$|\tilde{a}_{ij}(m)| = \left|\frac{a_{ij}(m)}{\sqrt{1+a_i}\sqrt{1+a_j}}\right| \leqslant |2a_{ij}(m)|,$$

we have

$$\begin{aligned} \sup_{i\geqslant 1, j\geqslant 1} P(\sqrt{m}|\tilde{a}_{ij}(m)| > t) &\leqslant \sup_{i\geqslant 1, j\geqslant 1} P(\sqrt{m}|2a_{ij}(m)| > t) \\ &= \sup_{i\geqslant 1, j\geqslant 1} P(\sqrt{m}|a_{ij}(m)| > t/2) \\ &= \exp(a - (b/4)t^2) + \frac{c}{\sqrt{m}} + \frac{4d}{t\sqrt{m}}. \end{aligned}$$

Hence, $\det(I_n + \frac{1}{n}\tilde{D}_{nm}) = 1 + o_p(1)$. □

Corollary 7.25. *Let* $I_n = diag(1 + a_1(m), \ldots, 1 + a_n(m))$ *and*

$$D_{nm} = \begin{pmatrix} a_{11}(m) & a_{12}(m) & \cdots & a_{1n}(m) \\ a_{21}(m) & a_{22}(m) & \cdots & a_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(m) & a_{n2}(m) & \cdots & a_{nn}(m) \end{pmatrix},$$

where $\forall t > 0$, $i \geqslant 1$, $j \geqslant 1$,

$$P(\sqrt{n_m}|a_{ij}(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$
$$P(\sqrt{n_m}|a_i(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$

for some constants a, b, c *and* d, *with* $b > 0$. *Then as* $m \to \infty$ *and* $n^4 = o(m)$, *we have*

$$[I_n + \frac{1}{n}D_{nm}]^{-1} = I_n + \frac{1}{n}\begin{pmatrix} b_{11}(m) & b_{12}(m) & \cdots & b_{1n}(m) \\ b_{21}(m) & b_{22}(m) & \cdots & b_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}(m) & b_{n2}(m) & \cdots & b_{nn}(m) \end{pmatrix},$$

where $\forall \epsilon > 0$, $\lim_{m\to\infty} \sup_{1 \leqslant i,j \leqslant n} P(|b_{ij}(m)| > \epsilon) = 0$.

Proof. The proof is tedious and omitted. The main idea is to use Cramer's rule and Lemma 7.24. □

Lemma 7.26. *Let* A *and* B *be* m *by* m *symmetric matrices. If* A *is invertible, and* $A - B$ *is invertible, then* $I_n \otimes A - \frac{1}{n}J_n \otimes B$ *is invertible and*

$$[I_n \otimes A - \frac{1}{n}J_n \otimes B]^{-1} = I_n \otimes A^{-1} + \frac{1}{n}J_n \otimes \{[A - B]^{-1} - A^{-1}\}$$

Proof.

$$[I_n \otimes A - \frac{1}{n}J_n \otimes B]^{-1} = (I_n \otimes A^{-\frac{1}{2}})[I - \frac{1}{n}J_n \otimes A^{-\frac{1}{2}}BA^{-\frac{1}{2}}]^{-1}(I_n \otimes A^{-\frac{1}{2}}).$$

Let $P = \frac{1}{n}A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$. Then,

$$\begin{aligned}
[I - \frac{1}{n}J_n \otimes A^{-\frac{1}{2}}BA^{-\frac{1}{2}}]^{-1} =& [I - J_n \otimes P]^{-1} \\
=& I + J_n \otimes P + (J_n \otimes P)^2 + (J_n \otimes P)^3 + \cdots \\
=& I + J_n \otimes P + nJ_n \otimes P^2 + n^2 J_n \otimes P^3 + \cdots \\
=& I + \frac{1}{n}J_n \otimes (-I_m + I_m + nP + (nP)^2 + \cdots) \\
=& I - \frac{1}{n}J_n \otimes \left\{[I_m - nP]^{-1} - I_m\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& [I_n \otimes A - \frac{1}{n}J_n \otimes B]^{-1} \\
=& (I_n \otimes A^{-\frac{1}{2}})[I - \frac{1}{n}J_n \otimes A^{-\frac{1}{2}}BA^{-\frac{1}{2}}]^{-1}(I_n \otimes A^{-\frac{1}{2}}) \\
=& (I_n \otimes A^{-\frac{1}{2}}) \left(I - \frac{1}{n}J_n \otimes \left\{[I_m - nP]^{-1} - I_m\right\}\right) (I_n \otimes A^{-\frac{1}{2}}) \\
=& I_n \otimes A^{-1} + \frac{1}{n}J_n \otimes \{[A - B]^{-1} - A^{-1}\}.
\end{aligned}$$

□

Corollary 7.27. *Suppose both* A *and* $A + B$ *are positive definite. Let*

$$X = \begin{pmatrix} A + a_1(m) & & \\ & \ddots & \\ & & A + a_n(m) \end{pmatrix} + \frac{1}{n} \begin{pmatrix} B + a_{11}(m) & B + a_{12}(m) & \cdots & B + a_{1n}(m) \\ B + a_{21}(m) & B + a_{22}(m) & \cdots & B + a_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ B + a_{n1}(m) & B + a_{n2}(m) & \cdots & B + a_{nn}(m) \end{pmatrix}.$$

Then, under the conditions of Lemma 7.24,

$$X^{-1} = \left[\begin{pmatrix} A & & \\ & \ddots & \\ & & A \end{pmatrix} + \frac{1}{n} \begin{pmatrix} B & B & \cdots & B \\ B & B & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & B \end{pmatrix} \right]^{-1} + \frac{1}{n} \begin{pmatrix} b_{11}(m) & b_{12}(m) & \cdots & b_{1n}(m) \\ b_{21}(m) & b_{22}(m) & \cdots & b_{2n}(m) \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}(m) & b_{n2}(m) & \cdots & b_{nn}(m) \end{pmatrix},$$

where $\forall t > 0$, $\lim_{m\to\infty} \max_{1\leqslant i,j\leqslant n} P(|b_{ij}(m)| > \epsilon) = 0$.

Proof. Without loss of generality, we consider the case in which A is an identity matrix. The following notations are introduced for convenience. Let

$$X = \begin{pmatrix} I + a_1 & & \\ & \ddots & \\ & & I + a_n \end{pmatrix} + \frac{1}{n} \begin{pmatrix} B + a_{11} & B + a_{12} & \cdots & B + a_{1n} \\ B + a_{21} & B + a_{22} & \cdots & B + a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B + a_{n1} & B + a_{n2} & \cdots & B + a_{nn} \end{pmatrix}$$

and

$$Y = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \frac{1}{n} \begin{pmatrix} B & B & \cdots & B \\ B & B & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & B \end{pmatrix}.$$

Then,

$$X = Y + \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix} + \frac{1}{n} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

By Lemma 7.26,

$$Y^{-1} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \frac{1}{n} \begin{pmatrix} C & C & \cdots & C \\ C & C & \cdots & C \\ \vdots & \vdots & \ddots & \vdots \\ C & C & \cdots & C \end{pmatrix}$$

for some C. Then,

$$Y^{-1}X = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix} + \frac{1}{n}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

$$+ \frac{1}{n}\begin{pmatrix} Ca_1 & Ca_2 & \cdots & Ca_n \\ Ca_1 & Ca_2 & \cdots & Ca_n \\ \vdots & \vdots & \ddots & \vdots \\ Ca_1 & Ca_2 & \cdots & Ca_n \end{pmatrix} + \frac{1}{n}\begin{pmatrix} C\bar{a}_{\cdot 1} & C\bar{a}_{\cdot 2} & \cdots & C\bar{a}_{\cdot n} \\ C\bar{a}_{\cdot 1} & C\bar{a}_{\cdot 2} & \cdots & C\bar{a}_{\cdot n} \\ \vdots & \vdots & \ddots & \vdots \\ C\bar{a}_{\cdot 1} & C\bar{a}_{\cdot 2} & \cdots & C\bar{a}_{\cdot n} \end{pmatrix}$$

$$= \begin{pmatrix} I + a_1 & & \\ & \ddots & \\ & & I + a_n \end{pmatrix} + \frac{1}{n}\begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \cdots & \tilde{a}_{1n} \\ \tilde{a}_{21} & \tilde{a}_{22} & \cdots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \cdots & \tilde{a}_{nn} \end{pmatrix}$$

$$=: Z$$

where $\bar{a}_{\cdot i} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$, $\tilde{a}_{ij} = a_{ij} + Ca_j + C\bar{a}_{\cdot j}$. We claim that if elementwisely

$$P(\sqrt{m}|a_{ij}| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{m}} + \frac{d}{t\sqrt{m}},$$

then there are constants $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$ with $\tilde{b} > 0$ such that elementwisely

$$P(\sqrt{m}|\tilde{a}_{ij}| > t) \leqslant \exp(\tilde{a} - \tilde{b}t^2) + \frac{\tilde{c}}{\sqrt{m}} + \frac{\tilde{d}}{t\sqrt{m}}.$$

By Corollary 7.25,

$$Z^{-1} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \frac{1}{n}\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix},$$

where $\forall \epsilon > 0$, as $m \to \infty$ and $n^4 = o(m)$, $\lim_{m\to\infty} \sup_{1\leqslant i,j\leqslant n} P(|b_{ij}| > \epsilon) = 0$. Therefore,

$$\begin{aligned} X^{-1} =& Z^{-1}Y^{-1} \\ =& Y^{-1} + \frac{1}{n}\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} Y^{-1} \\ =& Y^{-1} + \frac{1}{n}\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} + \frac{1}{n}\begin{pmatrix} \bar{b}_{1\cdot}C & \bar{b}_{1\cdot}C & \cdots & \bar{b}_{1\cdot}C \\ \bar{b}_{2\cdot}C & \bar{b}_{2\cdot} & \cdots & \bar{b}_{2\cdot} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{b}_{n\cdot}C & \bar{b}_{n\cdot} & \cdots & \bar{b}_{n\cdot}C \end{pmatrix} \\ =& Y^{-1} + \frac{1}{n}\begin{pmatrix} \tilde{b}_{11} & \tilde{b}_{12} & \cdots & \tilde{b}_{12} \\ \tilde{b}_{21} & \tilde{b}_{22} & \cdots & \tilde{b}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{b}_{n1} & \tilde{b}_{n2} & \cdots & \tilde{b}_{nn} \end{pmatrix}, \end{aligned}$$

where $\tilde{b}_{ij} = b_{ij} + \bar{b}_{i\cdot}C$. Since elementwisely $\lim_{m\to\infty} \sup_{1\leqslant i,j\leqslant n} P(|b_{ij}| > \epsilon) = 0$, it is easy to verify that elementwisely $\lim_{m\to\infty} \sup_{1\leqslant i,j\leqslant n} P(|\tilde{b}_{ij}| > \epsilon) = 0$. □

Lemma 7.28. *If*

- $f(t)$, $Ex(t)x(t)^T$ *and* $Ex(t)z(t)^T$ *have bounded derivatives*

- $\inf_t f(t) > 0$,
  $\inf_t \det(Ex(t)x(t)^T) > 0$,

- $E\sup_{0\leqslant t\leqslant T}[x(t)]^2 < \infty$,
  $E\sup_{0\leqslant t\leqslant T}[z(t)]^2 < \infty$,

- $C_{x,z}$ *and* $C_{z,z}$ *are continuous on* $[0,T]^2$, *and have bounded continuous second partial derivatives on* $\{(u,v): 0 \leqslant u < v \leqslant T\} \cup \{(u,v): 0 \leqslant v < u \leqslant T\}$,

then there exists a function g *increasing to infinity, such that* $\forall\, 0 < \delta < T/2$, *as* $n_s, n_m \to \infty$, $n_m = O(g(n_s))$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, *we have*

$$\begin{aligned}
&\frac{1}{n_s n_m} Z^T (I - S_d)^T R_d^{-1} (I - S_d) Z \\
&= \begin{pmatrix} A_d + a_1(n_m) & & & \\ & A_d + a_2(n_m) & & \\ & & \ddots & \\ & & & A_d + a_{n_c}(n_m) \end{pmatrix} \\
&\quad - \frac{1}{n_c} \begin{pmatrix} B_d + a_{11}(n_m) & B_d + a_{12}(n_m) & \cdots & \vdots \\ B_d + a_{21}(n_m) & B_d + a_{22}(n_m) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & B_d + a_{n_c n_c}(n_m) \end{pmatrix},
\end{aligned}$$

where $A_d = E\psi^d_{z,z}$, $B_d = E\psi^d_{z,z_{||x}} + E\psi^d_{z_{||x},z} - E\psi^d_{z_{||x},z_{||x}}$ *and* $\forall t > 0$, $i \geqslant 1$, $j \geqslant 1$,

$$P(\sqrt{n_m}|a_{ij}(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$
$$P(\sqrt{n_m}|a_i(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$

for some constants a, b, c *and* d, *with* $b > 0$.

Proof. Recall that $Z = \text{diag}(Z_1, \dots, Z_{n_c})$. By the definition of S_d, each row of $S_d Z$ can be expressed as

$$\begin{aligned}
& x_{ijk}^{\top}[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t_k, h_d)^{\top} Z \\
= & x_{ijk}^{\top}[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \underbrace{(\tilde{X}_1(t_k, h_d)^{\top} Z_1, \dots, \tilde{X}_{i'}(t_k, h_d)^{\top} Z_{i'}, \dots, \tilde{X}_{n_c}(t_k, h_d)^{\top} Z_{n_c})}_{n_c} \\
=: & \frac{1}{n_c} \underbrace{((\tilde{z}_{dijk}^{1})^{\top}, \dots, (\tilde{z}_{dijk}^{i'})^{\top}, \dots, (\tilde{z}_{dijk}^{n_c})^{\top})}_{n_c},
\end{aligned}$$

where

$$(\tilde{z}_{dijk}^{i'})^{\top} = x_{ijk}^{\top}[s_0(t_k, h_d) - s_1(t_k, h_d) s_2^{-1}(t_k, h_d) s_1(t_k, h_d)]^{-1} \frac{1}{n_s n_m} \tilde{X}_{i'}(t_k, h_d)^{\top} Z_{i'}.$$

Stack $(\tilde{z}_{dijk}^{i'})^{\top}$ to get

$$\tilde{Z}_{dij}^{i'} = \begin{pmatrix} (\tilde{z}_{dij1}^{i'})^{\top} \\ \vdots \\ (\tilde{z}_{dijn_m}^{i'})^{\top} \end{pmatrix}, \quad \tilde{Z}_{di}^{i'} = \begin{pmatrix} \tilde{Z}_{di1}^{i'} \\ \vdots \\ \tilde{Z}_{din_s}^{i'} \end{pmatrix}, \quad \tilde{Z}_{d}^{i'} = \begin{pmatrix} \tilde{Z}_{d1}^{i'} \\ \vdots \\ \tilde{Z}_{dn_c}^{i'} \end{pmatrix},$$

then we can write

$$S_d Z = \frac{1}{n_c} \begin{pmatrix} \tilde{Z}_{d1}^{1} & \tilde{Z}_{d1}^{2} & \cdots & \tilde{Z}_{d1}^{n_c} \\ \tilde{Z}_{d2}^{1} & \tilde{Z}_{d2}^{2} & \cdots & \tilde{Z}_{d2}^{n_c} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{Z}_{dn_c}^{1} & \tilde{Z}_{dn_c}^{2} & \cdots & \tilde{Z}_{dn_c}^{n_c} \end{pmatrix}.$$

Recall that $R_d = \text{diag}(R_{d1}, \dots, R_{dn_c})$, hence,

$$Z^{\top}(I - S_d)^{\top} R_d^{-1} (I - S_d) Z = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n_c} \\ A_{21} & A_{22} & \cdots & A_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n_c 1} & A_{n_c 2} & \cdots & A_{n_c n_c} \end{pmatrix},$$

where

$$A_{ii} = Z_i^\top R_{dc}^{-1} Z_i - \frac{1}{n_c} Z_i^\top R_{dc}^{-1} \tilde{Z}_{di}^i - \frac{1}{n_c} (\tilde{Z}_{di}^i)^\top R_{dc}^{-1} Z_i + \frac{1}{n_c^2} (\tilde{Z}_d^i)^\top R_d^{-1} \tilde{Z}_d^i,$$

and for $i \neq j$

$$A_{ij} = -\frac{1}{n_c} Z_i^\top R_{dc}^{-1} \tilde{Z}_{di}^j - \frac{1}{n_c} (\tilde{Z}_{dj}^i)^\top R_{dc}^{-1} Z_j + \frac{1}{n_c^2} (\tilde{Z}_d^i)^\top R_d^{-1} \tilde{Z}_d^j.$$

Define the following stochastic processes for $0 \leqslant t \leqslant T$,

$$\begin{aligned} z_{||x}^\top(t) =& x^\top(t)[\mathbf{E}(x(t)x^\top(t))]^{-1}\mathbf{E}(x(t)z(t)^\top), \\ z_{\perp x}^\top(t) =& z(t)^\top - z_{||x}^\top(t). \end{aligned}$$

We use the subscript "$||x$" and "$\perp x$", since

$$\mathbf{E}x(t)z_{||x}^\top(t) = \mathbf{E}x(t)z^\top(t), \quad \mathbf{E}x(t)z_{\perp x}^\top(t) = 0.$$

Recall $x_{ijk} = x_{ij}(t_k)$. Let

$$\begin{aligned} z_{ijk||x}^\top =& z_{||x_{ij}}^\top(t_k) = x_{ijk}^\top[\mathbf{E}(x(t_k)x^\top(t_k)|t_k)]^{-1}\mathbf{E}(x(t_k)z(t_k)^\top|t_k), \\ z_{ijk\perp x}^\top =& z_{ijk}^\top - z_{ijk||x}^\top. \end{aligned}$$

According to Corollary 7.19 and Lemma 7.21,

$$\begin{aligned} (\tilde{z}_{dijk}^{i'})^\top =& x_{ijk}^\top[s_0(t_k, h_d) - s_1(t_k, h_d)s_2^{-1}(t_k, h_d)s_1(t_k, h_d)]^{-1}\frac{1}{n_s n_m}\tilde{X}_{i'}(t_k, h_d)^\top Z_{i'} \\ =& x_{ijk}^\top\left[\frac{1}{f(t_k)}[\mathbf{E}x(t_k)x(t_k)^\top|t_k]^{-1} + o_p(1)\right]\left[f(t_k)\mathbf{E}[x(t_k)z(t_k)^\top|t_k] + o_p(1)\right] \\ =& x_{ijk}^\top([\mathbf{E}(x(t_k)x^\top(t_k)|t_k)]^{-1}\mathbf{E}(x(t_k)z(t_k)^\top|t_k) + o_p(1)) \\ =& z_{ijk||x}^\top + o_p(1). \end{aligned}$$

Stack $z_{ijk||x}^{\mathsf{T}}$ to get $Z_{i||x}$. According to Lemma 7.10,

$$
\begin{aligned}
\frac{1}{n_s} Z_i^{\mathsf{T}} R_{dc}^{-1} Z_i =& -\frac{1}{2\sigma_d^2 \log \rho_d} \sum_{k=1}^{n_m} \psi_{z,z}(t_k) + O_p(1), \\
\frac{1}{n_s} Z_i^{\mathsf{T}} R_{dc}^{-1} \tilde{Z}_{di}^j =& (1+o_p(1)) \frac{1}{n_s} Z_i^{\mathsf{T}} R_{dc}^{-1} Z_{i||x} \\
=& (1+o_p(1)) \left(-\frac{1}{2\sigma_d^2 \log \rho_d} \sum_{k=1}^{n_m} \psi_{z,z_{||x}}(t_k) + O_p(1) \right), \\
\frac{1}{n_s} (\tilde{Z}_{dj}^i)^{\mathsf{T}} R_{dc}^{-1} Z_j =& (1+o_p(1)) \frac{1}{n_s} Z_{j||x}^{\mathsf{T}} R_{dc}^{-1} Z_j \\
=& (1+o_p(1)) \left(-\frac{1}{2\sigma_d^2 \log \rho_d} \sum_{k=1}^{n_m} \psi_{z_{||x},z}(t_k) + O_p(1) \right), \\
\frac{1}{n_c n_s} (\tilde{Z}_d^i)^{\mathsf{T}} R_d^{-1} \tilde{Z}_d^j =& (1+o_p(1)) \frac{1}{n_c n_s} Z_{||x}^{\mathsf{T}} R_d^{-1} Z_{||x} \\
=& (1+o_p(1)) \left(-\frac{1}{2\sigma_d^2 \log \rho_d} \sum_{k=1}^{n_m} \psi_{z_{||x},z_{||x}}(t_k) + O_p(1) \right).
\end{aligned}
$$

Let $\psi_{\cdot,\cdot}^d(t) = -\frac{1}{2\sigma_d^2 \log \rho_d} \psi_{\cdot,\cdot}(t)$. Define

$$
\begin{aligned}
a_i =& \frac{1}{n_s n_m} Z_i^{\mathsf{T}} R_{dc}^{-1} Z_i - \mathbf{E}\psi_{z,z}^d, \\
a_{ij} =& \frac{1}{n_m} \left(\frac{1}{n_s} Z_i^{\mathsf{T}} R_{dc}^{-1} \tilde{Z}_{di}^j + \frac{1}{n_s} (\tilde{Z}_{dj}^i)^{\mathsf{T}} R_{dc}^{-1} Z_j - \frac{1}{n_c n_s} (\tilde{Z}_d^i)^{\mathsf{T}} R_d^{-1} \tilde{Z}_d^j \right) \\
& - (\mathbf{E}\psi_{z,z_{\perp x}}^d + \mathbf{E}\psi_{z_{\perp x},z}^d - \mathbf{E}\psi_{z_{\perp x},z_{\perp x}}^d),
\end{aligned}
$$

then,

$$
\begin{aligned}
\frac{1}{n_s n_m} A_{ii} =& \frac{1}{n_s n_m} \left(Z_i^{\mathsf{T}} R_{dc}^{-1} Z_i - \frac{1}{n_c} Z_i^{\mathsf{T}} R_{dc}^{-1} \tilde{Z}_{di}^i - \frac{1}{n_c} (\tilde{Z}_{di}^i)^{\mathsf{T}} R_{dc}^{-1} Z_i + \frac{1}{n_c^2} (\tilde{Z}_d^i)^{\mathsf{T}} R_d^{-1} \tilde{Z}_d^i \right) \\
=& \mathbf{E}\psi_{z,z}^d(t_1) + a_i - \frac{1}{n_c} \mathbf{E}\psi_{z,z_{||x}}^d(t_1) - \frac{1}{n_c} \mathbf{E}\psi_{z_{||x},z}^d(t_1) + \frac{1}{n_c} \mathbf{E}\psi_{z_{||x},z_{||x}}^d(t_1) + \frac{1}{n_c} a_{ij} \\
=:& A_d + a_i - \frac{1}{n_c}(B_d + a_{ij}),
\end{aligned}
$$

and for $i \neq j$

$$
\begin{aligned}
A_{ij} =& -\frac{1}{n_c} Z_i^T R_{dc}^{-1} \tilde{Z}_{di}^j - \frac{1}{n_c} (\tilde{Z}_{dj}^i)^T R_{dc}^{-1} Z_j + \frac{1}{n_c^2} (\tilde{Z}_d^i)^T R_d^{-1} \tilde{Z}_d^j \\
=& -\frac{1}{n_c} E\psi^d_{z,z_{||x}}(t_1) - \frac{1}{n_c} E\psi^d_{z_{||x},z}(t_1) + \frac{1}{n_c} E\psi^d_{z_{||x},z_{||x}}(t_1) + \frac{1}{n_c} a_{ij} \\
=:& -\frac{1}{n_c}(B_d + a_{ij}).
\end{aligned}
$$

Note that

$$\sqrt{n_m} a_i = n_m^{-1/2} \sum_{k=1}^{n_m} (\psi^d_{z,z}(t_k) - E\psi^d_{z,z}) + O_p(1),$$

and

$$
\begin{aligned}
\sqrt{n_m} a_{ij} =& (1 + o_p(1)) n_m^{-1/2} \Bigg(\sum_{k=1}^{n_m} \Big[\Big(\psi^d_{z,z_{||x}}(t_k) + \psi^d_{z_{||x},z}(t_k) - \psi^d_{z_{||x},z_{||x}}(t_k) \Big) - \\
& \Big(E\psi^d_{z,z_{\perp x}} + E\psi^d_{z_{\perp x},z} - E\psi^d_{z_{\perp x},z_{\perp x}} \Big) \Big] + O_p(1) \Bigg).
\end{aligned}
$$

Therefore by Lemma 7.23, $\exists a$, $b > 0$, c and d, elementwisely $\forall t > 0$, we have

$$
\begin{aligned}
P(\sqrt{n_m}|a_{ij}| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}}, \\
P(\sqrt{n_m}|a_i| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}}.
\end{aligned}
$$

□

Corollary 7.29. *If*

- $f(t)$, $Ex(t)x(t)^T$ *and* $Ex(t)z(t)^T$ *have bounded derivatives;*
- $\inf_t f(t) > 0$, $\inf_t \det(Ex(t)x(t)^T) > 0$;
- $E\sup_{0 \leqslant t \leqslant T}[x(t)]^2 < \infty$, $E\sup_{0 \leqslant t \leqslant T}[z(t)]^2 < \infty$;

- $C_{x,z}$ *and* $C_{z,z}$ *are continuous on* $[0,T]^2$*, and have bounded continuous second partial derivatives on* $\{(u,v) : 0 \leqslant u < v \leqslant T\} \cup \{(u,v) : 0 \leqslant v < u \leqslant T\}$*;*

- $E\psi^d_{z_{\perp x}, z_{\perp x}}$ *is invertible,*

then there exists a function g *increasing to infinity, such that* $\forall\, 0 < \delta < T/2$*, as* $n_s, n_m \to \infty$*,* $n_m = O(g(n_s))$*,* $n_c = O(n_s)$*,* $n_c^4 = o(n_m)$*,* $h_d \to 0+$*,* $1/h_d = O(\sqrt{n_m})$*,* $d = 1, 2$*, we have*

$$\left[\frac{1}{n_s n_m}\begin{pmatrix} Z^T(I-S_1)^T R_1^{-1}(I-S_1)Z & \\ & Z^T(I-S_2)^T R_2^{-1}(I-S_2)Z \end{pmatrix} + \frac{1}{n_s n_m} G^{-1}\right]^{-1}$$

$$= \begin{pmatrix} I_{n_c} \otimes [E\psi^1_{z,z}]^{-1} & \\ & I_{n_c} \otimes [E\psi^2_{z,z}]^{-1} \end{pmatrix}$$

$$+ \frac{1}{n_c}\begin{pmatrix} J_{n_c} \otimes ([E\psi^1_{z_{\perp x}, z_{\perp x}}]^{-1} - [E\psi^1_{z,z}]^{-1}) & \\ & J_{n_c} \otimes ([E\psi^2_{z_{\perp x}, z_{\perp x}}]^{-1} - [E\psi^2_{z,z}]^{-1}) \end{pmatrix}$$

$$+ \frac{1}{n_c}\begin{pmatrix} b_{11}(n_m) & \cdots & b_{1,2n_c}(n_m) \\ \vdots & \ddots & \vdots \\ b_{2n_c,1}(n_m) & \cdots & b_{2n_c,2n_c}(n_m) \end{pmatrix},$$

where $\forall \epsilon > 0$*, elementwisely* $\max_{1 \leqslant i,j \leqslant 2n_c} P(|b_{ij}(n_m)| > \epsilon) \to 0$*. We would like to note that the corollary is also valid when* G *is a general matrix as long as the absolute values of the elements of* G *are bounded. In particular, the corollary is valid when* G *is a zero matrix.*

Proof. Recall that G is $2n_c \times 2n_c$, and

$$G = \begin{pmatrix} D_{11} & & & D_{12} & & \\ & \ddots & & & \ddots & \\ & & D_{11} & & & D_{12} \\ D_{21} & & & D_{22} & & \\ & \ddots & & & \ddots & \\ & & D_{21} & & & D_{22} \end{pmatrix},$$

where

$$\begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = D = \mathrm{Var}(\beta_1) = \begin{pmatrix} \mathrm{Var}(\beta_{11}) & \mathrm{Cov}(\beta_{11}, \beta_{21}) \\ \mathrm{Cov}(\beta_{21}, \beta_{11}) & \mathrm{Var}(\beta_{21}) \end{pmatrix}.$$

Therefore,

$$G^{-1} = \begin{pmatrix} \tilde{D}_{11} & & & \tilde{D}_{12} & & \\ & \ddots & & & \ddots & \\ & & \tilde{D}_{11} & & & \tilde{D}_{12} \\ \tilde{D}_{21} & & & \tilde{D}_{22} & & \\ & \ddots & & & \ddots & \\ & & \tilde{D}_{21} & & & \tilde{D}_{22} \end{pmatrix},$$

where

$$\begin{pmatrix} \tilde{D}_{11} & \tilde{D}_{12} \\ \tilde{D}_{21} & \tilde{D}_{22} \end{pmatrix} = D^{-1}.$$

By Lemma 7.28,

$$\frac{1}{n_s n_m} Z^T (I - S_d)^T R_d^{-1} (I - S_d) Z$$

$$= \begin{pmatrix} A_d + a_1 & & & \\ & A_d + a_2 & & \\ & & \ddots & \\ & & & A_d + a_{n_c} \end{pmatrix}$$

$$- \frac{1}{n_c} \begin{pmatrix} B_d + a_{11} & B_d + a_{12} & \cdots & \vdots \\ B_d + a_{21} & B_d + a_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & B_d + a_{n_c n_c} \end{pmatrix},$$

where $A_d = \mathbf{E}\psi^d_{z,z}$, $B_d = \mathbf{E}\psi^d_{z,z_{||x}} + \mathbf{E}\psi^d_{z_{||x},x} - \mathbf{E}\psi^d_{z_{||x},z_{||x}}$, and $\forall t > 0$, $i \geqslant 1$, $j \geqslant 1$,

$$P(\sqrt{n_m}|a_{ij}(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$
$$P(\sqrt{n_m}|a_i(n_m)| > t) \leqslant \exp(a - bt^2) + \frac{c}{\sqrt{n_m}} + \frac{d}{t\sqrt{n_m}},$$

for some constants a, $b > 0$, c and d. Further,

$$\frac{1}{n_s n_m} G^{-1} = \frac{1}{n_c} \begin{pmatrix} \frac{n_c}{n_s n_m} O_{11} & \frac{n_c}{n_s n_m} O_{12} & \cdots & \vdots \\ \frac{n_c}{n_s n_m} O_{21} & \frac{n_c}{n_s n_m} O_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \frac{n_c}{n_s n_m} O_{2n_c, 2n_c} \end{pmatrix},$$

where O_{ij} is $\tilde{D}_{11}$, or $\tilde{D}_{12}$, or $\tilde{D}_{21}$, or $\tilde{D}_{22}$ or a all-zero matrix. So

$$\frac{1}{n_s n_m} \left[\begin{pmatrix} Z^T(I - S_1)^T R_1^{-1}(I - S_1)Z & \\ & Z^T(I - S_2)^T R_2^{-1}(I - S_2)Z \end{pmatrix} + G^{-1} \right]$$

$$= \begin{pmatrix} I_{n_c} \otimes A_1 & \\ & I_{n_c} \otimes A_2 \end{pmatrix} + \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{2n_c} \end{pmatrix}$$

$$- \frac{1}{n_c} \begin{pmatrix} J_{n_c} \otimes B_1 & \\ & J_{n_c} \otimes B_2 \end{pmatrix} - \frac{1}{n_c} \begin{pmatrix} a_{11} - \frac{n_c}{n_s n_m} O_{11} & \cdots & a_{2n_c,1} - \frac{n_c}{n_s n_m} O_{2n_c,1} \\ \vdots & \ddots & \vdots \\ a_{2n_c,1} - \frac{n_c}{n_s n_m} O_{2n_c,1} & \cdots & a_{2n_c,2n_c} - \frac{n_c}{n_s n_m} O_{2n_c,2n_c} \end{pmatrix},$$

where a_{ij} is a q by q zero matrix if not previously specified. If n_c/n_s is bounded, then we have

$$P(\sqrt{n_m}|a_{ij}(n_m) - \frac{n_c}{n_s n_m} O_{ij}| > t) \leqslant \exp(\tilde{a} - \tilde{b}t^2) + \frac{\tilde{c}}{\sqrt{n_m}} + \frac{\tilde{d}}{t\sqrt{n_m}},$$

for some constants $\tilde{a}$, $\tilde{b} > 0$, $\tilde{c}$, and $\tilde{d}$. Thus we can apply Corollary 7.27. According to Lemma 7.28, Corollary 7.27 and Lemma 7.26,

$$\left[\frac{1}{n_s n_m} \begin{pmatrix} Z^T (I - S_1)^T R_1^{-1} (I - S_1) Z & \\ & Z^T (I - S_2)^T R_2^{-1} (I - S_2) Z \end{pmatrix} + \frac{1}{n_s n_m} G^{-1} \right]^{-1}$$

$$= \begin{pmatrix} I_{n_c} \otimes A_1^{-1} & \\ & I_{n_c} \otimes A_2^{-1} \end{pmatrix}$$

$$+ \frac{1}{n_c} \begin{pmatrix} J_{n_c} \otimes ([A_1 - B_1]^{-1} - A_1^{-1}) & \\ & J_{n_c} \otimes ([A_2 - B_2]^{-1} - A_2^{-1}) \end{pmatrix}$$

$$+ \frac{1}{n_c} \begin{pmatrix} b_{11} & \cdots & b_{1,2n_c} \\ \vdots & \ddots & \vdots \\ b_{2n_c,1} & \cdots & b_{2n_c,2n_c} \end{pmatrix},$$

where $A_d = \mathbf{E}\psi^d_{z,z}$, $B_d = \mathbf{E}\psi^d_{z,z_{||x}} + \mathbf{E}\psi^d_{z_{||x},x} - \mathbf{E}\psi^d_{z_{||x},z_{||x}}$. And $\forall \epsilon > 0$,

$$\lim_{m\to\infty} \max_{1\leqslant i,j\leqslant 2n_c} P(|b_{ij}(m)| > \epsilon) = 0.$$

Note that $A_d - B_d = \mathbf{E}\psi^d_{z_{\perp x}, z_{\perp x}}$ is positive definite. □

Asymptotic property of $\tilde{X}(t, h_d)m_d$

Lemma 7.30. *Let*

$$\begin{aligned}\Delta(t,h) =& \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k - t) x_{ijk} x_{ijk}^T \alpha''_d(\tilde{t}_k)(t_k - t)^r \\ & - h^r \mu_r f(t) \mathbf{E}[x(t)x(t)^T] \alpha''_d(t),\end{aligned}$$

where $\tilde{t}_k$ *lies between* t_k *and* t*, and satisfies* $\alpha''_d(t_k) = \alpha_d(t) + \alpha'_d(t)(t_k - t) + \frac{1}{2}\alpha''_d(\tilde{t}_k)(t_k - t)^2$*. If* f *has a bounded derivative and* $\mathbf{E}x(t)x(t)^T$ *as a function of* t *has a bounded derivative, then* $\forall\, 0 < \delta < T/2$*, as* $n_m \to \infty$*,* $h \to 0+$*,* $1/h = O(\sqrt{n_m})$*, and* $n_c n_s \to \infty$*, elementwisely we have*

$$\sup_{\delta\leqslant t\leqslant T-\delta} |\Delta(t,h)| = o_p(h^r).$$

Proof. Elementwisely

$$\begin{aligned}&|\Delta(t,h)| \\ \leqslant& |s_r(t,h)\alpha''_d(t) - \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k - t) x_{ijk} x_{ijk}^T \alpha''_d(\tilde{t}_k)(t_k - t)^r| \\ &+ |s_r(t,h)\alpha''_d(t) - h^r \mu_r f(t) \mathbf{E}[x(t)x(t)^T] \alpha''_d(t)|.\end{aligned}$$

We shall prove that

$$(1) \sup_{\delta\leqslant t\leqslant T-\delta} |s_r(t,h)\alpha''_d(t) - \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k - t) x_{ijk} x_{ijk}^T \alpha''_d(\tilde{t}_k)(t_k - t)^r| = o_p(h^r),$$

(2) $\sup_{\delta \leqslant t \leqslant T-\delta} |s_r(t,h)\alpha_d''(t) - h^r \mu_r f(t) \mathbf{E}[x(t)x(t)^T]\alpha_d''(t)| = o_p(h^r)$.

For (1), elementwisely let $\delta(h) = \sup_{|u-v|\leqslant h} |\alpha_d''(u) - \alpha_d''(v)|$. Note that $\delta(h)$ is a p by 1 vector. Since α_d'' is continuous on a closed interval, it is also uniformly continuous, which implies that $\lim_{h\to 0_+} \delta(h) = 0$. Elementwisely

$$\begin{aligned}
&\left| s_r(t,h)\alpha_d''(t) - \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k - t) x_{ijk} x_{ijk}^T \alpha_d''(\tilde{t}_k)(t_k - t)^r \right| \\
=&\left| \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k - t) x_{ijk} x_{ijk}^T \big(\alpha_d''(\tilde{t}_k) - \alpha_d''(t)\big)(t_k - t)^r \right| \\
\leqslant& \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \left| K_h(t_k - t) x_{ijk} x_{ijk}^T \big(\alpha_d''(\tilde{t}_k) - \alpha_d''(t)\big)(t_k - t)^r \right| \\
\leqslant& \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \left| K_h(t_k - t) x_{ijk} x_{ijk}^T (t_k - t)^r \right| |\alpha_d''(\tilde{t}_k) - \alpha_d''(t)| \\
\leqslant& \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \left| K_h(t_k - t) x_{ijk} x_{ijk}^T (t_k - t)^r \right| \delta(h).
\end{aligned}$$

Similar to the proof of Lemma 7.15, we can show that

$$\sup_{\delta \leqslant t \leqslant T-\delta} \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \left| K_h(t_k - t) x_{ijk} x_{ijk}^T (t_k - t)^r \right| = O(h^r),$$

so as $h \to 0$,

$$\sup_{\delta \leqslant t \leqslant T-\delta} \frac{1}{n_c n_s n_m} \sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} \left| K_h(t_k - t) x_{ijk} x_{ijk}^T (t_k - t)^r \right| \delta(h) = \delta(h) O_p(h^r) = o_p(h^r).$$

For (2), by Lemma 7.15, if f has a bounded derivative and $\mathbf{E}x(t)x(t)^T$ as a function of t has a bounded derivative, then $\forall\, 0 < \delta < T/2$, as $n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, and $n_c n_s \to \infty$, elementwisely we have

$$\sup_{\delta \leqslant t \leqslant T-\delta} \left| \frac{s_r(t,h) - h^r \mu_r f(t) \mathbf{E}x(t)x(t)^T}{h^r} \right| = o_p(1).$$

That is, $\sup_{\delta\leqslant t\leqslant T-\delta}|s_r(t,h)\alpha_d''(t)-h^r\mu_r f(t)\mathbf{E}[x(t)x(t)^\top]\alpha_d''(t)|=o_p(h^r)$. □

Lemma 7.31. *Let* $\Delta(t,h)$ *be the difference,*

$$\left([s_0(t,h)-s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1}\frac{1}{n_c n_s n_m}\tilde{X}(t,h)^\top m_d\right)-\left(\alpha_d(t)+\frac{1}{2}h^2\mu_2\alpha_d''(t)\right).$$

Then if f *has a bounded derivative,* $\mathbf{E}x(t)x(t)^\top$ *as a function of* t *has a bounded derivative, and* $\inf_t f(t)>0$, $\inf_t \det(\mathbf{E}x(t)x(t)^\top)>0$, *then* $\forall\, 0<\delta<T/2$, *as* $n_m\to\infty$, $h\to 0+$, $1/h=O(\sqrt{n_m})$, *and* $n_c n_s\to\infty$, *elementwisely*

$$\sup_{\delta\leqslant t\leqslant T-\delta}|\Delta(t,h)|=o_p(h^2).$$

Proof.

$$\begin{aligned}
&\frac{1}{n_c n_s n_m}\tilde{X}(t,h)^\top m_d\\
=&\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}(I_p-s_1s_2^{-1}(t_k-t))K_h(t_k-t)x_{ijk}x_{ijk}^\top\alpha_d(t_k)\\
=&\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}(I_p-s_1s_2^{-1}(t_k-t))K_h(t_k-t)x_{ijk}x_{ijk}^\top\left(\alpha_d(t)+\alpha_d'(t)(t_k-t)\right)\\
&+\frac{1}{2}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}(I_p-s_1s_2^{-1}(t_k-t))K_h(t_k-t)x_{ijk}x_{ijk}^\top\alpha_d''(\tilde{t}_k)(t_k-t)^2)\\
=&(s_0-s_1s_2^{-1}s_1)\alpha_d(t)+(s_1-s_1s_2^{-1}s_2)\alpha_d'(t)\\
&+\frac{1}{2}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}K_h(t_k-t)x_{ijk}x_{ijk}^\top\alpha_d''(\tilde{t}_k)(t_k-t)^2\\
&-s_1s_2^{-1}\frac{1}{2}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}K_h(t_k-t)x_{ijk}x_{ijk}^\top\alpha_d''(\tilde{t}_k)(t_k-t)^3\\
=&(s_0-s_1s_2^{-1}s_1)\alpha_d(t)\\
&+\frac{1}{2}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}K_h(t_k-t)x_{ijk}x_{ijk}^\top\alpha_d''(\tilde{t}_k)(t_k-t)^2
\end{aligned}$$

$$-s_1 s_2^{-1}\frac{1}{2}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k-t)x_{ijk}x_{ijk}^{\mathsf T}\alpha_d''(\tilde t_k)(t_k-t)^3, \qquad (*)$$

where $\tilde t_k$ lies between t_k and t. By Lemma 7.30, if f has a bounded derivative and $\mathbf{E}x(t)x(t)^{\mathsf T}$ as a function of t has a bounded derivative, then $\forall\, 0 < \delta < T/2$, as $n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, and $n_c n_s \to \infty$, we have

$$\sup_{\delta \leqslant t \leqslant T-\delta}\left|\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m} K_h(t_k-t)x_{ijk}x_{ijk}^{\mathsf T}\alpha_d''(\tilde t_k)(t_k-t)^r - h^r\mu_r f(t)\mathbf{E}[x(t)x(t)^{\mathsf T}]\alpha_d''(t)\right| = o_p(h^r),$$

Therefore,

$$\begin{aligned}(*) =&(s_0 - s_1 s_2^{-1}s_1)\alpha_d(t)\\ &+\frac{1}{2}(h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^{\mathsf T}]\alpha_d''(t) + o_p(h^2))\\ &-\frac{1}{2}s_1 s_2^{-1}(h^3\mu_3 f(t)\mathbf{E}[x(t)x(t)^{\mathsf T}]\alpha_d''(t) + o_p(h^3)). \qquad (**)\end{aligned}$$

By Lemma 7.15, elementwisely we have

$$\sup_{\delta \leqslant t \leqslant T-\delta}\left|\frac{s_r(t,h) - h^r\mu_r f(t)\mathbf{E}x(t)x(t)^{\mathsf T}}{h^r}\right| = o_p(1).$$

By Corollary 7.18, if $\inf_t \det(\mathbf{E}x(t)x(t)^{\mathsf T}) > 0$, then

$$\sup_{\delta \leqslant t \leqslant T-\delta}\left|h^2[s_2(t,h)]^{-1} - \frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^{\mathsf T}]^{-1}\right| = o_p(1),$$

Therefore,

$$\begin{aligned}(**) =&(s_0 - s_1 s_2^{-1}s_1)\alpha_d(t)\\ &+\frac{1}{2}(h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^{\mathsf T}]\alpha_d''(t) + o_p(h^2))\\ &-\frac{1}{2}s_1(h^2 s_2^{-1})(h\mu_3 f(t)\mathbf{E}[x(t)x(t)^{\mathsf T}]\alpha_d''(t) + o_p(h))\\ =&(s_0 - s_1 s_2^{-1}s_1)\alpha_d(t)\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}(h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2))\\
&-\frac{1}{2}\Big[h\mu_1 f(t)\mathbf{E}[x(t)x(t)^T]+ho_p(1)\Big]\Big[\frac{1}{\mu_2 f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}+o_p(1)\Big]\times\\
&\quad\times\Big[h\mu_3 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h)\Big]\\
=&(s_0-s_1s_2^{-1}s_1)\alpha_d(t)\\
&+\frac{1}{2}(h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2))\\
&-\frac{1}{2}o_p(h^2)\left[\frac{1}{\mu_2 f(t)}(\mathbf{E}[x(t)x(t)^T])^{-1}+o_p(1)\right]\\
=&(s_0-s_1s_2^{-1}s_1)\alpha_d(t)+\frac{1}{2}(h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2))-o_p(h^2)\\
=&(s_0-s_1s_2^{-1}s_1)\alpha_d(t)+\frac{1}{2}h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&[s_0-s_1s_2^{-1}s_1]^{-1}\frac{1}{n_c n_s n_m}\sum_{i=1,j=1,k=1}^{n_c,n_s,n_m}(I_p-s_1s_2^{-1}(t_k-t))K_h(t_k-t)x_{ijk}x_{ijk}^T\alpha_d(t_k)\\
=&[s_0-s_1s_2^{-1}s_1]^{-1}\left[[s_0-s_1s_2^{-1}s_1)]\alpha_d(t)+\frac{1}{2}h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2)\right]\\
=&\alpha_d(t)+[s_0-s_1s_2^{-1}s_1]^{-1}\left[\frac{1}{2}h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2)\right].
\end{aligned}
$$

By Corollary 7.19,

$$
\sup_{\delta\leqslant t\leqslant T-\delta}\left|[s_0(t,h)-s_1(t,h)s_2^{-1}(t,h)s_1(t,h)]^{-1}-\frac{1}{f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}\right|=o_p(1),
$$

Hence,

$$
\begin{aligned}
&\alpha_d(t)+[s_0-s_1s_2^{-1}s_1]^{-1}\left[\frac{1}{2}h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2)\right]\\
=&\alpha_d(t)+\frac{1}{f(t)}[\mathbf{E}x(t)x(t)^T]^{-1}\left[\frac{1}{2}h^2\mu_2 f(t)\mathbf{E}[x(t)x(t)^T]\alpha_d''(t)+o_p(h^2)\right]\\
=&\alpha_d(t)+\frac{1}{2}h^2\mu_2\alpha_d''(t)+o_p(h^2).
\end{aligned}
$$

□

Asymptotic property of $Z^T(I-S_d)^T R_d^{-1}(I-S_d)m_d$

Now consider $Z^T(I-S_d)^T R_d^{-1}(I-S_d)m_d$.

Lemma 7.32. *There exists a function* g *increasing to infinity such that as* $n_s, n_m \to \infty$, $n_m = O(g(n_s))$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, *we have* $Z^T(I-S_d)^T R_d^{-1}(I-S_d)m_d(A_1^T \cdots A_{n_c}^T)^T$, *where elementwisely,*

$$\frac{1}{n_s n_m} A_i = h^2(E\psi^d_{z_{\perp x}, u_d} + o_p(1)).$$

Proof. Recall that each row of S_d is $x_{ijk}^T[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t_k, h_d)^T$, and by Lemma 7.31, as $n_m \to \infty$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, and $n_c n_s \to \infty$, elementwisely,

$$\sup_{\delta \leqslant t \leqslant T-\delta} |[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t,h)^T m_d - \alpha_d(t) - \frac{1}{2} h^2 \mu_2 \alpha_d''(t)| = o_p(h^2).$$

Note that each row of m_d is $x_{ijk}^T \alpha_d(t_k)$, so each row of $(I-S_d)m_d = m_d - S_d m_d$ is of the form

$$x_{ijk}^T \alpha_d(t_k) - x_{ijk}^T(\alpha_d(t) + \frac{1}{2} h^2 \mu_2 \alpha_d''(t) + o_p(h^2)) = h^2(-\frac{1}{2}\mu_2 x_{ijk}^T \alpha_d''(t_k) + o_p(1)).$$

Define $u_d(t) = -\frac{1}{2}\mu_2 x(t)^T \alpha_d''(t)$ and let $u_{dijk} = -\frac{1}{2}\mu_2 x_{ijk}^T \alpha''(t_k)$. Stack u_{dijk} to get U_{dij}, U_{di} and U_d. In the recall of Lemma 7.28, we write

$$S_d Z = \frac{1}{n_c} \begin{pmatrix} \tilde{Z}_{d1}^1 & \tilde{Z}_{d1}^2 & \cdots & \tilde{Z}_{d1}^{n_c} \\ \tilde{Z}_{d2}^1 & \tilde{Z}_{d2}^2 & \cdots & \tilde{Z}_{d2}^{n_c} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{Z}_{dn_c}^1 & \tilde{Z}_{dn_c}^2 & \cdots & \tilde{Z}_{dn_c}^{n_c}. \end{pmatrix}$$

Therefore,

$$\begin{aligned}
&Z^T(I-S_d)^T R_d^{-1}(I-S_d)m_d\\
=&Z^T(I-S_d)^T R_d^{-1}(U_d+o_p(1))\\
=&Z^T R_d^{-1}(U_d+o_p(1))+(S_dZ)^T R_d^{-1}(U_d+o_p(1))\\
=&(A_1^T\cdots A_{n_c}^T)^T,
\end{aligned}$$

where $A_i = h^2\left(Z_i^T R_{dc}^{-1} U_{di} - \frac{1}{n_c}(\tilde{Z}_d^i)^T R_d^{-1} U_d + o_p(1)\right)$. According to Lemma 7.10, there exists a function g increasing to infinity such that as $n_s \to \infty$ and $n_m = O(g(n_s))$,

$$\begin{aligned}
\frac{1}{n_s} Z_i^T R_{dc}^{-1} U_{di} =& -\frac{1}{2\sigma_d^2 \log \rho_d}\sum_{k=1}^{n_m}\psi_{z,u}(t_k)+O_p(1),\\
\frac{1}{n_s n_c}(\tilde{Z}_d^i)^T R_d^{-1} U_d =& \frac{1}{n_c}\sum_{i'=1}^{n_c}\frac{1}{n_s}(\tilde{Z}_{di'}^i)^T R_{dc}^{-1} U_{di'}\\
=&\frac{1}{n_c}\sum_{i'=1}^{n_c}\left(-\frac{1}{2\sigma_d^2 \log \rho_d}\sum_{k=1}^{n_m}\psi_{z_{||x},u}(t_k)+O_p(1)\right)\\
=&-\frac{1}{2\sigma_d^2 \log \rho_d}\sum_{k=1}^{n_m}\psi_{z_{||x},u}(t_k)+O_p(1).
\end{aligned}$$

Let $\psi^d_{\cdot,\cdot}(t) = -\frac{1}{2\sigma_d^2 \log \rho_d}\psi_{\cdot,\cdot}(t)$. Since

$$\begin{aligned}
\frac{1}{n_s n_m} Z_i^T R_{dc}^{-1} U_{di} =& \mathbf{E}\psi^d_{z,u_d} + O_p(\frac{1}{\sqrt{n_m}}),\\
\frac{1}{n_c n_s n_m}(\tilde{Z}_d^i)^T R_d^{-1} U_d =& \mathbf{E}\psi^d_{z_{||x},u_d} + O_p(\frac{1}{\sqrt{n_m}}),
\end{aligned}$$

we have

$$\frac{1}{n_s n_m} A_i = h^2(\mathbf{E}\psi^d_{z_{\perp x},u_d} + o_p(1)).$$

□

We would like to note that through the proof of Lemma 7.10, $\forall \epsilon > 0$, as $n_s, n_m \to \infty$, $n_m = O(g(n_s))$, $h \to 0+$, $1/h = O(\sqrt{n_m})$, elementwisely

$$\max_{1 \leqslant i \leqslant n_c} P\left(\left|\frac{1}{n_s n_m h^2} A_i - E\psi^d_{z_{\perp x}, u_d}\right| > \epsilon\right) \to 0.$$

Asymptotic property of $\tilde{X}(t, h_d)E_d$

Lemma 7.33. *If* $E \sup_t |x_{ij}(t)| < \infty$, $E \sup_t |\epsilon_{ij}(t)| < \infty$, *and* $f(t)$ *has a bounded derivative, then under the conditions in Corollary 7.19, Lemma 7.15, and Corollary 7.18, there is a function* g *increasing to infinity such that*

$$\sup_t \left|[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t, h_d)^T E_d\right| = o_p\left(\frac{1}{g(n_c n_s)}\right).$$

Proof. Let $w_r(t, h) = \frac{1}{n_c n_s n_m} \sum_{i,j,k} K_h(t_k - t) x_{ijk} \epsilon_{ijk} (t_k - t)^r$, then we can write

$$[s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t, h_d)^T E_d = [s_0 - s_1 s_2^{-1} s_1]^{-1} [w_0 - s_1 s_2^{-1} w_1].$$

We first have

$$\begin{aligned} |w_r(t, h)| &= \left|\frac{1}{n_m} \sum_k K_h(t_k - t)(t_k - t)^r \frac{\sum_{i,j} x_{ijk}\epsilon_{ijk}}{n_c n_s}\right| \\ &\leqslant \frac{1}{n_m} \sum_k K_h(t_k - t)|t_k - t|^r \left|\frac{\sum_{i,j} x_{ijk}\epsilon_{ijk}}{n_c n_s}\right| \\ &\leqslant \sup_t \left|\frac{\sum_{i,j} x_{ij}(t)\epsilon_{ij}(t)}{n_c n_s}\right| h^r \frac{1}{n_m} \sum_k K_h(t_k - t). \end{aligned}$$

In the second step, by Lemma 7.14, $\sup_t |\frac{1}{n_m} \sum_k K_h(t_k - t)| = O_p(1)$, hence

$$\sup_t |w_r(t, h)| = h^r o_p\left(\frac{1}{g(n_c n_s)}\right).$$

By Corollary 7.19, Lemma 7.15, and Corollary 7.18,

$$\sup_t \left|[s_0 - s_1 s_2^{-1} s_1]^{-1}[w_0 - s_1 s_2^{-1} w_1]\right| = o_p(\frac{1}{g(n_c n_s)}).$$

□

Asymptotic property of $Z^T(I - S_d)^T R_d^{-1}(I - S_d)E_d$

Now consider $Z^T(I - S_d)^T R_d^{-1}(I - S_d)E_d$. The analysis here is quite similar to the analysis of Lemma 7.28, so we will not go into details as we did for Lemma 7.28.

Lemma 7.34. *Under the conditions of Lemma 7.33 and 7.10, we have*

$$\frac{1}{n_s n_m} C_i = O_p(1/n_m).$$

Proof. By the definition of S_d, each row of $S_d E_d$ can be expressed as

$$x_{ijk}^T [s_0 - s_1 s_2^{-1} s_1]^{-1} \frac{1}{n_c n_s n_m} \tilde{X}(t_k, h_d)^T E_d.$$

Applying Lemma 7.33 to each row of $S_d E_d$, we get $S_d E_d = o_p(1/g(n_c n_s))$. Then,

$$Z^T(I - S_d)^T R_d^{-1}(I - S_d)E_d = (C_1^T \cdots C_{n_c}^T)^T + o_p(1),$$

where

$$\begin{aligned} C_i =& Z_i^T R_{dc}^{-1} E_{i\perp x} - \frac{1}{n_c}(\tilde{Z}_d^i)^T R_d^{-1} E_{\perp x} \\ =& Z_i^T R_{dc}^{-1} E_{i\perp x} - \frac{1}{n_c} \sum_{i'=1}^{n_c} (\tilde{Z}_{di'}^i)^T R_{dc}^{-1} E_{i'\perp x}. \end{aligned}$$

According to Lemma 7.10,

$$\frac{1}{n_s} Z_i^T R_{dc}^{-1} E_{i\perp x} = \sum_{k=1}^{n_m} \psi_{z,\epsilon_{d\perp x}}^d (t_k) + O_p(1),$$

$$\frac{1}{n_s}(\tilde{Z}^i_{di'})^\top R^{-1}_{dc} E_{i'|x} = \sum_{k=1}^{n_m} \psi^d_{z_{||x},\epsilon_{d\perp x}}(t_k) + O_p(1),$$

$$\frac{1}{n_c}\sum_{i'=1}^{n_c}\frac{1}{n_s}(\tilde{Z}^i_{di'})^\top R^{-1}_{dc} E_{i'|x} = \sum_{k=1}^{n_m} \psi^d_{z_{||x},\epsilon_{d\perp x}}(t_k) + O_p(1).$$

By the definition of ψ, we have $\psi^d_{z,\epsilon_{d\perp x}} = \psi^d_{z_{||x},\epsilon_{d\perp x}} = 0$. Therefore,

$$\frac{1}{n_s n_m} C_i = O_p(1/n_m).$$

□

Asymptotic variance of $\hat{\alpha}_d$

According to Lemma 7.34, the principal term of $(I - S_d)E_d$ is E_d, so

$$\begin{aligned}
&M^6_d(M^6_d)^\top \\
\approx&\tilde{X}^\top Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top(I-S_d)^\top R_d^{-1}E_d \\
&\times E_d^\top R_d^{-1}(I-S_d)Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X} \\
\approx&\tilde{X}^\top Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top(I-S_d)^\top R_d^{-1}R_d \\
&\times R_d^{-1}(I-S_d)Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X} \\
=&\tilde{X}^\top Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X} \\
\approx&n_c n_s n_m f^2(t)\mathbf{E}[x(t)z^\top(t)][\mathbf{E}\psi^d_{z_{\perp x},z_{\perp x}}]^{-1}\mathbf{E}[z(t)x^\top(t)].
\end{aligned}$$

$$\begin{aligned}
&M^3_d(M^6_d)^\top \\
\approx&\tilde{X}^\top E_d E_d^\top R_d^{-1}(I-S_d)Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X} \\
\approx&\tilde{X}^\top R_d R_d^{-1}(I-S_d)Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X} \\
=&\tilde{X}^\top(I-S_d)Z[Z^\top(I-S_d)^\top R_d^{-1}(I-S_d)Z]^{-1}Z^\top\tilde{X}
\end{aligned}$$

$$\approx \tilde{X}^{\mathsf{T}} Z_{\perp x}[Z^{\mathsf{T}}(I - S_d)^{\mathsf{T}} R_d^{-1}(I - S_d)Z]^{-1} Z^{\mathsf{T}} \tilde{X}$$

$$\approx 0.$$

$$\begin{aligned}
M_d^3(M_d^3)^{\mathsf{T}} =& \tilde{X}^{\mathsf{T}} E_d E_d^{\mathsf{T}} \tilde{X} \\
=& \sum_{i,j,k,i',j',k'} [I_p - s_1 s_2^{-1}(t_k - t)] K_{h_d}(t_k - t) x_{ijk} \epsilon_{ijk} \\
& \times \left([I_p - s_1 s_2^{-1}(t_{k'} - t)] K_{h_d}(t_{k'} - t) x_{i'j'k'} \epsilon_{i'j'k'}\right)^{\mathsf{T}} \\
\approx & \sum_{i,j,k,i',j',k'} K_{h_d}(t_k - t) K_{h_d}(t_{k'} - t) x_{ijk} x_{i'j'k'}^{\mathsf{T}} \epsilon_{ijk} \epsilon_{i'j'k'} \\
\approx & n_c n_s \sum_{k,k'} K_{h_d}(t_k - t) K_{h_d}(t_{k'} - t) A(t_k, t_{k'}) \\
\approx & n_c n_s \sum_{k,k'} K_{h_d}(t_k - t) K_{h_d}(t_{k'} - t) A(t, t)
\end{aligned}$$

where $A(t_k, t_{k'}) = \mathbf{E}[x_{ijk} x_{ijk'}^{\mathsf{T}} \epsilon_{ijk} \epsilon_{ijk'} | t_k, t_{k'}]$. By Lemma 7.14, $(n_m)^{-2} \sum_{k,k'} K_{h_d}(t_k - t) K_{h_d}(t_{k'} - t) = f^2(t) + o_p(1)$. So

$$M_d^3(M_d^3)^{\mathsf{T}} \approx n_m^2 n_c n_s f^2(t) \mathbf{E}[x(t) x^{\mathsf{T}}(t) \epsilon^2(t)].$$

The asymptotic variance of $\hat{\alpha}_d$ is determined by $[s_{0d} - s_{1d} s_{2d}^{-1} s_{1d}]^{-1} \frac{1}{n_c n_s n_m} M_d^3$ and $[s_{0d} - s_{1d} s_{2d}^{-1} s_{1d}]^{-1} \frac{1}{n_c n_s n_m} M_d^6$ (see Subsection 7.3.2). Since

$$[s_{0d} - s_{1d} s_{2d}^{-1} s_{1d}]^{-1} \approx [f^2(t) \mathbf{E} x(t) x^{\mathsf{T}}(t)]^{-1},$$

the asymptotic variance of $\hat{\alpha}_d$ is $\Omega V \Omega$, where

$$\Omega = [\mathbf{E} x(t) x^{\mathsf{T}}(t)]^{-1}, \quad V = \mathbf{E}[x(t) x^{\mathsf{T}}(t) \epsilon^2(t)].$$

and the convergence rate is $\sqrt{n_c n_s}$.

Chapter 8

Bibliography

Ahsanullah, M., and Nevzorov, V. B. (2015). Records via Probability Theory. Atlantis Press.

Song, X. Y., Cai, J. H., Feng, X. N., & Jiang, X. J. (2014). Bayesian analysis of the functional-coefficient autoregressive heteroscedastic model. Bayesian Analysis, 9(2), 371-396.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. Transactions of the american mathematical society, 49(1), 122-136.

Bickel, P. and Levian, E. (2006). Regularized estimation of large covariance matrices. The Annals of Statistics, 36, 199-227.

Bronnenberg, B.J. and Mahajan, V. (2001). Unobserved Retailer Behavior in Multi-market Data: Joint Shares and Promotion Variables. Marketing Science, 20, 284-299.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. Journal of Econometrics, 136(1), 163-188.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for non-linear time series. Journal of the American Statistical Association, 95(451), 941-956.

Chan, W. S., and Tse, Y. K. (1993). Price-volume relation in stocks: A multiple time series analysis on the Singapore market. Asia Pacific Journal of Management, 10(1), 39-56.

Chen, K. and Jin, Z. (2005). Local polynomial regression analysis of clustered data. Biometrika, 92, 59-74.

Chi, E.M. and Reinsel G.C. (1989). Models for Longitudinal Data With Random Effects and AR(1) Errors. Journal of the American Statistical Association, 84, 452-459.

Cochrane, J. H. (2009). Asset Pricing:(Revised Edition). Princeton university press.

Coval, J., Jurek, J., and Stafford, E. (2009). The economics of structured finance. The Journal of Economic Perspectives, 23(1), 3-25.

Creala, D., Koopmanb, S. J., and Lucasc, A. (2011). The estimation of time-varying parameters in multivariate linear time series models.

Daniels, M. J., and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. Biometrics, 57(4), 1173-1184.

Delsing, M. J., Oud, J. H., and De Bruyn, E. E. (2005). Assessment of bidirectional influences between family relationships and adolescent problem behavior. European Journal of Psychological Assessment, 21(4), 226-231.

Demidenko, E. (2004). Mixed Models: Theory and Applications. Wiley Press.

Deschenes, O., Greenstone, M. (2007). Climate change, mortality and adaption: Evidence from annual fluctuations in weather in the U.S. Working Paper, the National Bureau of Economic Research.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). Analysis of longitudinal data, 2nd ed. Oxford University Press.

Doob, J. L. (1942). The Brownian movement and stochastic equations. Annals of Mathematics, 351-369.

Engle, R., Granger, C.W., Rice J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association, 81, 310-320.

Esseen, C. G. (1942). On the Liapounoff limit of error in the theory of probability. Almqvist and Wiksell.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Journal of the American Statistical Association, 99, 710-723.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. Journal of the American Statistical Association, 103, 1520-1533.

Fan, J. and Zhang J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society: Series B, 62, 303-322.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. The Annals of Statistics, 27, 1491-1518.

Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. The Annals of Statistics, 2008-2036.

Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. Journal of the American Statistical Association, 102, 632-641.

Fan. J. and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications. London: Chapman & Hall.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(3), 369-390.

Frees, EW. (1995). Assessing cross-sectional correlation in panel data. Journal of Econometrics, 69, 393-414.

Frees, EW. (2001). Omitted variables in longitudinal data models. The Canadian Journal of Statistics, 29, 573-595.

Frees, EW. (2004). Longitudinal and panel data: analysis and applications in the social sciences. Cambridge University Press.

Functional coefficient autoregressive models for vector time series

Generalized Linear Mixed Models with Varying Coefficients for Longitudinal Data

Granger, C. W., and Morgenstern, O. (1963). SPECTRAL ANALYSIS OF NEW YORK STOCK MARKET PRICES1. Kyklos, 16(1), 1-27.

Greene, W. (2002). Econometric Analysis, 5th edition. Prentice Hall.

Hackl, P., & Westlund, A. H. (1995). On price elasticities of international telecommunication demand. Information Economics and Policy, 7(1), 27-36.

Hackl, P., & Westlund, A. H. (1996). Demand for international telecommunication time-varying price elasticity. Journal of Econometrics, 70(1), 243-260.

Hall, P., Horowitz, J. L., & Jing, B. Y. (1995). On blocking rules for the bootstrap with dependent data. Biometrika, 82(3), 561-574.

Hamao, Y., Masulis, R. W., and Ng, V. (1990). Correlations in price changes and volatility across international stock markets. Review of Financial studies, 3(2), 281-307.

Hansen, E. Lecture note on week convergence. Retrieved January 23, 2015, from http://www.math.ku.dk/ erhansen/WeakConv13/doku/noter/All.pdf

Hassan, M. R., and Nath, B. (2005). Stock market forecasting using hidden Markov model: a new approach. Intelligent Systems Design and Applications, 5th International Conference, 192-196.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society: Series B, 55, 757-796.

Heagerty, P. and Zeger S. (2000). Marginalized multilevel models and likelihood inference. Statistics Science, 15, 1-26.

Hull, J. C. (2006). Options, futures, and other derivatives. Pearson Education India.

Jiang, X. Q. (1999). Time varying coefficient AR and VAR models. In The Practice of Time Series Analysis (pp. 175-191). Springer New York.

Kapoor, M., Kelejian, H.H. and Prucha, I.R. (2007). Panel data models with spatially correlated error components. Journal of Econometrics, 140, 97-130.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. Journal of Financial and quantitative Analysis, 22(01), 109-126.

Kelejian, H.H. and Prucha, I.R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. Journal of Real Estate Finance and Economics, 17, 99-121.

Kelejian, H.H. and Prucha, I.R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. International Economic Review, 40, 509-533.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 963-974.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generailized linear models. Biometrika, 73, 13-22.

Lin, X. and Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. Journal of the American Statistical Association, 95, 520-534.

Lin, X. and Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations. Journal of the American Statistical Association, 96, 1045-1056.

Lin, X. and Carroll, R.J. (2006). Semiparametric estimation in general repeated measuers problems. Journal of the Royal Statistical Society: Series B, 68, 69-88.

Lindstrom, M., and Bates, D. (1988). Newton-Raphson and EM Algorithms for Linear Miexed-Effects Models for Repeated-Measures Data. Journal of the American

Statistical Association, 83, 1014-1022.

Liu, L., Ma J. and O'Quigley J. (2008). Joint analysis of multi-level repeated measures data and survival: an application to the end stage renal disease (ESRD) data. Statistics in Medicine, 27, 5679-5691.

Lu, F., Qiao, H., Wang, S., Lai, K. K., and Li, Y. (2016). Time-varying coefficient vector autoregressions model based on dynamic correlation with an application to crude oil and stock markets. Environmental Research.

Mack, Y. P., & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. Zeitschrift fÃ¼r Wahrscheinlichkeitstheorie und verwandte Gebiete, 61(3), 405-415.

Magnus, J.R. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. Journal of Econometrics, 19, 239-285.

MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data

Olsen, M., Delong, E., Oddone E. and Bosworth H. (2008) Strategies for analyzing multilevel cluster-randomized studies with binary outcomes collected at varying intervals of time. Statistics in Medicine, 27, 6055-6071.

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2009). A hierarchical OrnsteinâŁ"Uhlenbeck model for continuous repeated measurement data. Psychometrika, 74(3), 395-418.

Orbe, S., Ferreira, E., & Rodriguez-Poo, J. (2006). On the estimation and testing of time varying constraints in econometric models. Statistica Sinica, 1313-1333.

Pai, P. F., and Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6), 497-505.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies, 72(3), 821-852.

R Chen Functional-Coefficient Autoregressive Models

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. Journal of Applied Statistics, 37(9), 1487-1498.

Rice, J.A., and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the dare are curves. Journal of the Royal Statistical Society: Series B, 53, 233-243.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In Statistical Analysis and Forecasting of Economic Structural Change (pp. 253-264).

Springer Berlin Heidelberg.

Ruppert, D., Sherther, S.J. and Wand, M.P. (1995). And effective bandwidth selector for local least square regression. Journal of the American Statistical Association, 90, 1257-1270.

Shao, J. (2003). Mathematical Statistics, Springer-Verlag.

Sun, W. (2003). Relationship between trading volume and security prices and returns. Area Exam Report, Technical Report, MIT Laboratory for Information and Decision Systems.

Sun, Y., Zhang, W. and Tong, H. (2007). Estimation of the Covariance Matrix of Random Effects in Longitudinal Studies. The Annals of Statistics, 35, 2795-2814.

Swamy, P., Chang, I., Mehta, J. and Tavlas, G. (2003). Correcting for omitted-variable and measurement-error bias in autogressive model estimation with panel data. Computational Economics, 22, 225-253.

Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., and Wong, T. L. (2007). *Design and implementation of NN5 for Hong Kong stock price forecasting*. Engineering Applications of Artificial Intelligence, 20(4), 453-461.

Tsay, R. S. (2005). Analysis of financial time series (Vol. 543). John Wiley & Sons.

Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

Verbeke, G. and Molenberghts, G. (2000). Linear Mixed Models for Longitudinal Data, Springer-Verlag.

Wand, M. P., and Jones, M. C. (1994). Kernel smoothing. Crc Press.

Wang, N. (2003). Marginal nonparametric kernal regression accounting for within-subject correlation. Biometrika, 90, 43-52.

Wang, N., Carrol, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. Journal of the American Statistical Association, 100, 147-157.

Wang, Y., and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association, 96(455), 895-905.

White, H. (1994). Estimation, Inference and Specification Analysis, New York: Cambridge Uniersity Press.

Woodbury, M. A. (1950). Inverting modified matrices. Memorandum report, 42, 106.

Wooldridge, J. (2001). Econometric Analysis of Cross Section and Panel Data, The MIT Press.

Wu, B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika, 90, 831-844.

Wu, H. and Zhang, J. (2002). Local polynomial mixed-effects models for longitudinal data. Journal of the American Statistical Association, 97, 833-897.

Zimmer, D.M., 2012, The role of copulas in the housing crisis, Review of Economics and Statistics, 94, 607-620.