The Pennsylvania State University
The Graduate School

NEW MODELS FOR CONDITIONAL COVARIANCE MATRIX

A Dissertation in

Statistics

by

Ying Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2013

The dissertation of Ying Zhang was reviewed and approved* by the following:

Runze Li
Professor of Statistics
Dissertation Advisor, Chair of Committee

Bing Li
Professor of Statistics

Zhibiao Zhao
Assistant Professor of Statistics

Qiang Du
Verne M. Willaman Professor of Mathematics

David Hunter
Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

This thesis presents new models for conditional covariance matrix. The proposed nonparametric covariance regression model parameterizes the conditional covariance matrix of a multivariate response vector as a quadratic function of regression splines. The resulting conditional covariance matrix is positive definite for all explanatory variables and represents the conditional covariance as the summation of a "baseline" covariance matrix and a positive definite matrix depending on the explanatory variables. The proposed approach provides an adaptable representation of heteroscedasticity across the levels of explanatory variable. In addition, the model has a random-effect representation, allowing for the maximum likelihood parameter estimation via the EM-algorithm. The asymptotic normality for the estimators is established and some numerical examples are used to illustrate the proposed procedure.

To cope with the high-dimensionality of the covariates, estimating the conditional covariance matrix through a modified Cholesky decomposition is proposed. The modified Cholesky decomposition procedure associates each local covariance matrix with a unique unit lower triangular and a unique diagonal matrix. The entries of the lower triangular matrix and the diagonal matrix have statistical interpretation as regression coefficients and prediction variances when regressing each term on its predecessors. It ensures that the estimated conditional covariance matrix is positive definite. To circumvent the curse of dimensionality, a class of partially linear models are used to estimate those regression coefficients and local linear estimators are developed to estimate the nonparametric variance functions. The asymptotic properties of the proposed procedure are studied. We show that the proposed procedure for estimating the conditional covariance matrix based on residuals has the same asymptotic bias and variance as that based on true errors. Comprehensive simulation studies and a real data example are presented to illustrated the proposed methods.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply indebted to my advisors Dr. Runze Li for his guidance, encouragement and mentorship throughout my graduate studies. This work would not have been possible without his support both intellectually and financially. Moreover, I would like to thank Dr. Bing Li, Dr. Zhibiao Zhao, and Dr. Qiang Du for serving on my committee and providing valuable suggestions to improve my dissertation.

This dissertation research has been supported by National Science Foundation grant DMS 0348869, a grant of National Institute on Drug Abuse, NIH, P50-DA10075 and a grant of National Institute of Cancer, NIH, R01 CA168676.

I would like to thank my parents for their love and support along the way. Finally I would like to thank my husband, Yicheng Wen, for his patience, love and support through this almost never-ending process.

Chapter 1

Introduction

1.1 Motivation of the Research

The problem of mean regression (i.e., $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{E}[\mathbf{y}|\mathbf{x}]$) has been well studied in both the univariate and multivariate settings. However, estimating a conditional covariance function $\Sigma_{\mathbf{x}} = \mathrm{Var}[\mathbf{y}|\mathbf{x}]$ across a range of response values for an explanatory $\mathbf{x}$-variable is less studied. In the univariate case, a number of statistical models and procedures suggest that the variance can be expressed as a function of the mean, i.e. for some known function g, $\sigma^2_{\mathbf{x}} = g(\mu_{\mathbf{x}})$. For example, see the discussion in Carroll et al. (1982). Other approaches include separately estimating the mean and the variance (see, for example, Rutemiller & Bowers (1968), Smyth (1989), etc.) and using kernel estimates of the variance function (Müller & Stadtmüller (1987)).

The multivariate covariance regression has generally been developed by standard regression operations on the unconstrained elements of the logarithm of the Cholesky decomposition of the covariance or the precision matrix. Chiu et al. (1996b) suggested modelling the elements of the logarithm of the covariance matrix (i.e., $\Phi_{\mathbf{x}} = \log\Sigma_{\mathbf{x}}$) as linear functions of the explanatory variables, i.e. $\phi_{j,k,x} = \beta^T_{j,k}\mathbf{x}$ for unknown coefficients $\beta_{j,k}$. One advantage of the constructed covariance function $\{\Sigma_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$ is that the domain of the explanatory x-variable is the same as that in mean regression, i.e., the explanatory vector can be continuous, discrete, and categorical. However, an issue with this formulation is the difficulty of the parameter interpretation; a submatrix of $\Sigma_{\mathbf{x}}$ does not necessarily coincide with a submatrix of $\Phi_{\mathbf{x}}$, so the elements of $\Phi_{\mathbf{x}}$ do not directly relate to the corresponding covariances in $\Sigma_{\mathbf{x}}$. Pourahmadi (1999) proposed to model

the unconstrained elements of the Cholesky decomposition of $\Sigma_{\mathbf{x}}^{-1}$ as linear functions of $\mathbf{x}$. The weights associated with the jth row have a natural interpretation in terms of the conditional distribution of y_j given $y_1,\ldots,y_{j-1}$. However, one problem of such method is that this model is not invariant to reordering of the elements of $\mathbf{y}$, and therefore, it is problematic when there is no natural order of the variables. Additionally, both the models of Chiu et al. (1996b) and Pourahmadi (1999) require $q\times p(p+1)/2$ parameters to be estimated, which is very large. The idea of Cholesky decomposition has also been used by Pourahmadi (2000), Wu & Pourahmadi (2003), and Huang et al. (2006) to reformulate the estimation of covariance matrix within the framework of regression modeling.

In recent years, data sets with high dimension and small sample size relative to dimension have become very common. Examples include gene expression arrays, spectroscopic imaging, numerical weather forecasting, and many others. Estimating large covariance matrices, where the dimension of the data p is comparable to or larger than the sample size n, has gained particular attention latterly, since high-dimensional data are so common in applications. There have been a series of researches focusing on the estimation of sparse covariance matrix or precision matrix using regularization procedures. For example, see the discussions by Meinshausen & Bühlmann (2006), Bickel & Levina (2008a,b), Levina et al. (2008), and Lam & Fan (2009). Meinshausen & Bühlmann (2006) proposed a computationally attractive method for covariance selection that can be used for sparse high-dimensional graphs by performing neighborhood selection with the LASSO for each node in the graph. By either banding or tapering the sample covariance matrix, or estimating a banded version of precision matrix, Bickel & Levina (2008b) obtained an estimate of large covariance matrix which were shown to be consistent in the operator normal if $(\log p)/n\to 0$. When the variables in covariance matrix have a natural ordering, Levina et al. (2008) used the Cholesky decomposition and introduced a nested Lasso penalty to estimate the large covariance matrix. Lam & Fan (2009) precisely studied the sparsistency and the rate of convergence for estimating sparse covariance and precision matrices based on penalized likelihood with nonconvex penalty functions.

Nonparametric regression models have been commonly used in assorted areas. A number of researchers have extensively studied the various estimation procedures for the nonparametric regression function. To our best knowledge, although there are some

references on nonparametric conditional variance function (see Ruppert et al. (1997), Fan & Yao (1998) and references therein), references for nonparametric models for conditional covariance matrix are very limited. Lately, Fan et al. (2007) introduced to model the correlation functions using parametric models and the variance functions using a fully nonparametric model to estimate the conditional covariance functions of response variable conditioning on a set of covariates. A kernel estimator for estimating the nonparametric variance function was developed and the positive definiteness of the estimated covariance function was ensured. Yin et al. (2010) proposed a fully nonparametric model for the conditional covariance matrix. However, one limitation of this method is that it is restricted to the high-dimensional covariates.

More recently, Hoff & Niu (2012) proposed a covariance regression model that parameterizes the covariance matrix of a vector of multivariate response as a parsimonious quadratic function of explanatory variables, i.e., $\Sigma_{\mathbf{x}} = \mathbf{A} + \mathbf{B}\mathbf{x}\mathbf{x}^T\mathbf{B}^T$ with $\mathbf{A}$ positive definite and $\mathbf{B}$ real. The $q \times p$ parameters of $\mathbf{B}$ have a direct interpretation in terms of how heteroscedasticity co-occurs among the p variables of $\mathbf{y}$. In addition, the model has a random-effects representation, so that straightforward maximum likelihood parameter estimation via the EM-algorithm can be used, which is computationally efficient. However, this model still has some limitations in: (1) flexibility based on the parametric approach; (2) high-dimensional case with large p . In addition, Hoff & Niu (2012) have not proven any theoretical properties of the estimated parameters.

1.2 Contribution of the Thesis

In this thesis, I consider nonparametric, conditional covariance functions. The proposed covariance regression model parameterizes the conditional covariance matrix of a multivariate response vector as a quadratic function of regression splines, which can be regarded as a natural extension of Hoff & Niu (2012)'s method. The resulting covariance function is positive definite for all explanatory variables and represents the conditional covariance as a "baseline" covariance matrix plus a positive definite matrix depending on the explanatory variables. The proposed approach provides an adaptable representation of heteroscedasticity across the levels of explanatory variable. The explanatory variables of all types, including categorical variables, are accommodated in the proposed covariance regression model, which is useful in the analysis of multivariate data. The

maximum likelihood parameter estimation via the EM-algorithm is used for estimating the parameters. The theoretical development has also been established, which shows the asymptotic normality for the estimators. Several numerical examples are used to illustrate the proposed procedure.

The proposed nonparametric conditional covariance models result only considers the case where the covariates are in the low dimensional space. However, it becomes less useful in situations where the covariates are high-dimensional. Therefore, estimating the conditional covariance matrix through a modified Cholesky decomposition is proposed. The modified Cholesky decomposition procedure associates each local covariance matrix with a unique unit lower triangular and a unique diagonal matrix. The entries of the lower triangular matrix and the diagonal matrix have statistical interpretation as regression coefficients and prediction variances when regressing each term on its predecessors. To circumvent the curse of dimensionality of covariates, a class of partially linear models are used to estimate those regression coefficients and kernel estimators are developed to estimate the nonparametric covariance functions. This proposed method ensures that the estimated conditional covariance function is positive definite locally. It also retains the parsimony of parametric models and flexibility of the nonparametric models. The asymptotic properties of the proposed procedure are studied. Comprehensive simulation studies and a real data example are presented to illustrated the proposed methods.

1.3 Organization of the Thesis

The rest of this thesis is divided into the following chapters.

- Chapter 2 provides a brief review of the related work regarding the estimation of the covariance matrix. First, parameterization-based approaches, including spectral decomposition, variance-correlation decomposition, and Cholesky decomposition are discussed. Second, several regularization procedures for estimating large covariance matrices, where the dimension of the data is comparable to or larger than the sample size, are presented. Two broad groups of covariance estimations are considered: those that rely on a natural ordering among variables, assuming that variables far apart in the ordering are only weakly correlated and those invariant under variable permutations. Third, nonparametric approaches for estimating the covariance structures are discussed.

- Chapter 3 proposes a general approach to estimate the conditional covariance matrix in the context of nonparametric regression models. Two covariance regression models are considered: one with continuous explanatory variable only and the other with both continuous and discrete explanatory variables. Parameter estimation for each model is made via random-effects representation and an EM-algorithm. The consistency and asymptotic normality properties of the maximum likelihood estimators (MLEs) for the proposed nonparametric covariance regression model are shown. A number of simulation studies are also examined in this chapter, with an application to the Boston Housing data presented.

- Chapter 4 proposes a methodology for estimating the conditional covariance matrix in the context of high dimensionality. A modified Cholesky decomposition procedure associates each local covariance matrix with a unique unit lower triangular and a unique diagonal matrix. The entries of the lower triangular matrix and the diagonal matrix have statistical interpretation as regression coefficients and prediction variances when regressing each term on its predecessors. A class of partially linear models are used to estimate those regression coefficients and kernel estimators are developed to estimate the nonparametric variance functions. The asymptotic properties of the proposed procedure are studied. Comprehensive simulation studies and a real data example are presented to illustrated the proposed methods.

- Chapter 5 concludes the dissertation with recommendation of the future work on the proposed nonparametric models for the conditional covariance matrix.

Chapter 2

Literature Review

The estimation of covariance matrix is one of the most common and important tasks in statistical analysis. It has profound applications in assorted fields, which include but are not limited to: graphical modeling (see, for example, Edwards (2000), Drton & Perlman (2004), Yuan & Lin (2007)), longitudinal data analysis (see, for example, Diggle & Verbyla (1998), Smith & Kohn (2002)), machine learning (see, for example, Bilmes (2000)), and multivariate volatility in finance (see, for example, Bollerslev et al. (1988), Engle (2002)), etc. It is also one the most challenging and difficult task in practice due to its dimensionality and the positive definite constraint. In this chapter, several major methods for the estimation of covariance matrix are presented.

2.1 Parameterization-based Approaches

In consideration of the complexity of a covariance matrix, it is helpful to start by breaking it down into components based on modelling considerations and mathematical convenience. Decompositions that have caught research interests include spectral decomposition, variance-correlation decomposition, and Cholesky decomposition. All these methods consist of decomposing complicated covariance matrices into "dependence" and "variance" components and then modeling them virtually and separately using regression techniques (Pourahmadi et al. (2007)).

2.1.1 Spectral (Eigenvalue) Decomposition

The spectral decomposition (also known as eigenvalue decomposition) is to reparameterize a covariance matrix Σ_k in terms of its eigenvalue and eigenvector:

$$\Sigma_k = D_k \Lambda_k D_k^T, \tag{2.1}$$

where D_k is the matrix of eigenvectors and Λ_k is a diagonal matrix with the eigenvalues of Σ_k on the diagonal.

The reparameterization of covariance matrices in terms of the eigenvalue decomposition has been considered by Flury (1984, 1988). In that model, Flury assumed the eigenvector matrices, D_k, to be the same across all population and thus the covariance matrix in the kth population can be expressed as

$$\Sigma_k = D \Lambda_k D^T. \tag{2.2}$$

If samples are from independent multivariate normal populations and the eigenvectors are uniquely identified except for sign and permutation, the maximum likelihood estimates of D and Λ_k can be computed by the Flury-Gautschi algorithm (Flury & Constantine (1985); Flury & Gautschi (1986); Clarkson (1988)) and the asymptotic distribution of the estimators is known (Flury (1986a)).

In the context of Gaussian clustering, Banfield & Raftery (1993) have used the spectral decomposition to specify that some, but not all, features (orientation, size, or shape) to be the same for all clusters. They considered a reparameterization of the covariance matrix Σ_k of a cluster P_k in terms of its eigenvalue decomposition,

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{2.3}$$

where λ_k defines the volume of P_k, D_k is an orthogonal matrix which determines its orientation, and A_k is a diagonal matrix with determinant 1 which defines its shape. They also gave algorithms for computing the maximum likelihood estimates, but did not obtain the asymptotic distribution of the estimators. The proportional covariance model (Owen (1984); Flury (1986b); Eriksen (1987); Manly & Rayner (1987); Schott (1999)) is a special case of Banfield & Raftery (1993)'s model. Celeux & Govaert (1995) have considered the eigenvalue decomposition of the clusters' covariance matrices from a

general and flexible point of view. They proposed many general clustering criteria from the simplest one (spherical clusters with equal volumes which leads to the classical k-means criterion) to the most complex one (unknown and different volumes, orientations and shapes for all clusters). To overcome the limitations of Banfield & Raftery (1993)'s approach, which includes no assessment of the uncertainty about the classification, bias for the estimated parameter, pre-specified shape matrix by user, equal prior group probabilities, and no formal way of choosing among the possible models etc., Bensmail et al. (1997) proposed a fully Bayesian analysis of the model based clustering methodology of Banfield & Raftery (1993). Boik (2002) proposed a spectral model for the simultaneous eigenstructure of multiple covariance matrices, which subsumes most existing common principal components and related models. Under normality, he used a Fisher scoring algorithm for computing maximum likelihood estimates of the parameters and derived the asymptotic distributions of the estimators.

In spectral decomposition (2.1), the matrix D_k can be further decomposed into a product of Givens rotation matrices, so that Σ_k is parameterized in terms of its eigenvalues and Givens angles. (see, e.g., Yang & Berger (1994); Pinheiro & Bates (1996); Daniels & Kass (1999); and references therein). Yang & Berger (1994) placed a reference prior on the eigenvalues and the Givens angles and then used it to carry out Bayesian inference on the covariance matrix. Under different loss functions, the performance of the resulting Bayes estimators of the covariance matrix was showed to be comparable with several alternative estimators. The approach of Yang & Berger (1994) is flexible in that it does not assume any specific parametric form for the covariance matrix, and is applied generally to covariance matrices arising from cross-sectional data as well as those from longitudinal data. However, it does not identify any specific parsimonious structure in the covariance matrix, which is an desirable objective for the covariance matrix under some specific settings, for instance, longitudinal data.

In transforming to the matrix logarithm, Leonard & Hsu (1992) substantially started from the spectral decomposition. They proposed a flexible class of prior distribution for the covariance matrix of a multivariate normal distribution and developed exact and approximate Bayesian, empirical and hierarchical Bayesian estimation and finite sample inference techniques. Applying the spectral decomposition, Chiu et al. (1996a) provided a flexible methodology to model the structure of a covariance matrix and study the dependence of the covariances on explanatory variables. They suggested modelling the

elements of the logarithm of the covariance matrix,

$$\Psi_x = \log \Sigma_x, \tag{2.4}$$

as linear functions of the explanatory variables, so that for unknown coefficients $\beta_{i,j}$,

$$\psi_{i,j,x} = \beta_{i,j}^T \mathbf{x}. \tag{2.5}$$

This generalized linear model for covariance matrices provides an applicable addition to the existing range of special structures for covariance matrices. The proposed method makes use of the fact that the only constraint on Ψ_x is its symmetry. However, as the authors noted, parameter interpretation for the model is difficult. For instance, a submatrix of Σ_x is not generally the matrix exponential of the same submatrix of Ψ_x, so the elements of Ψ_x do not directly link to the corresponding covariances in Σ_x. On the other hand, the number of parameters in the model can be quite large. For $\mathbf{y} \in \mathbb{R}^p$ and $\mathbf{x} \in \mathbb{R}^q$, the model involves a q-dimensional vector of coefficients for each of the $p(p+1)/2$ unique elements of Ψ_x, so the total parameters to be estimated is $q \times p(p+1)/2$.

2.1.2 Variance-Correlation Decomposition

Another commonly used method for handling the covariance matrix is based on the variance-correlation decomposition. Manly & Rayner (1987) introduced a hierarchy and a corresponding ANOVA-type partition of the likelihood ratio test statistic for the comparison of two or more sample covariance matrices. They showed that the differences between the covariance matrices depends on changing variances, changing correlations, and matrices being proportional. Barnard et al. (2000) modeled a covariance matrix in terms of its corresponding standard deviations and correlation matrix. Specifically, they wrote

$$\Sigma_k = P_k R_k P_k^T,$$

where $P_k = \text{diag}(\sqrt{\sigma_{k11}}, \ldots, \sqrt{\sigma_{kpp}})$ is a diagonal matrix whose diagonal entries are the square-roots of those of Σ_k, and R_k is the corresponding $p \times p$ correlation matrix. Two general modelling situations where the variance-correlation decomposition approach is applicable and useful: shrinkage estimation of regression coefficients and a general location-scale model for both continuous and categorical variables, were dis-

cussed. It is noticed that the positive definite constraint of the covariance matrix can be easily handled via a Gibbs-sampler formulation.

2.1.3 Cholesky Decomposition

In addition, the idea of Cholesky decomposition is widely used to estimate the unconditional covariance matrix in recent literature. Liu (1993) used the Cholesky decomposition to obtain a Bartlett-type decomposition of the posterior distribution of a covariance matrix with monotone missing data. There is also a literature on using the Cholesky decomposition directly for the covariance matrix (Pinheiro & Bates (1996)), though the resulting parameterizations do not have simple statistical interpretation. Pourahmadi (1999, 2000) used the Cholesky decomposition of the inverse of a covariance matrix to associate a unique unit lower triangular and a unique diagonal matrix with each covariance matrix. The covariance parameters have statistical interpretation as the regression coefficients and logarithms of prediction error variances corresponding to regressing a response on its predecessor. More specifically, the modified Cholesky decomposition can be viewed as:

$$T\Sigma T^T = V, \tag{2.6}$$

where T is a unit lower triangular matrix, which has unconstrained entries with statistical interpretation as the generalized autoregressive parameters (GARP), and the entries of $V = \text{diag}(v_1^2, \ldots, v_p^2)$ are the corresponding residual variance (Pourahmadi (1999)). More concretely, let $Y = (Y_1, \ldots, Y_p)$ be a generic random vector with mean zero and positive-definite covariance matrix Σ. Let $\hat{Y}_i$ stand for the linear least-squares predictor of Y_i based on its predecessors $Y_{i-1}, \ldots, Y_1$ and ε_i be its prediction error:

$$\begin{aligned} \hat{Y}_i &= \sum_{j=1}^{i-1} \phi_{i,j} Y_j, \\ \varepsilon_i &= Y_i - \hat{Y}_i = Y_i - \sum_{j=1}^{i-1} \phi_{i,j} Y_j, \; i = 1, \ldots, p, \end{aligned} \tag{2.7}$$

where the regression coefficients $\phi_{i,j}$'s are unconstrained and the variances $v_i^2 = \text{Var}(\varepsilon_i)$ are non-negative. Evidently, the prediction errors are uncorrelated, so that, with $\varepsilon =$

$(\varepsilon_1,\ldots,\varepsilon_p)^T$, it follows that $\mathrm{Cov}(\varepsilon) = \mathrm{diag}(v_1^2,\ldots,v_p^2) = V$. Writing (2.7) in matrix form one obtains

$$\varepsilon = TY, \tag{2.8}$$

where T is a unit lower triangular matrix with $-\phi_{i,j}$ in the (i,j)th position for $2 \leq i \leq p$ and $j = 1,\ldots,i-1$. From (2.8), we can obtain that the matrix T diagonalizes the covariance matrix Σ as in (2.6). This diagonalization is related to the modified Cholesky decomposition of Σ and Σ^{-1} . It is clear that this decomposition depends on the ordering of the components of Y, so it is well suited to data that have ordered responses, such as longitudinal data. In a follow-up paper (Pourahmadi (2000)), the maximum likelihood estimators of the parameters of a generalized linear model for the covariance matrix, their consistency and their asymptotic normality were studied under the condition that the observations are normally distributed.

In the analysis of longitudinal data, parsimonious modelling of the covariance structure is of great importance. Smith & Kohn (2002) proposed a data-driven approach to identify parsimony in the covariance matrix of longitudinal data. A statistically efficient estimator of the covariance matrix was obtained by factoring the inverse of the covariance matrix using the Cholesky decomposition. Their method differs from Pourahamadi's in that parsimony was built into the model by allowing the elements in the strict lower triangle matrix to be identically equal to zero; however, Pourahmadi did not attempt to formally identify any structural zeros. A hierarchical Bayesian model was used to flexibly identify any such zeros. The model was estimated using a Markov chain Monte Carlo (MCMC) sampling scheme that is computationally efficient and can be applied to covariance matrices of high dimension. In a follow-up paper to Pourahmadi (1999, 2000), Daniels & Pourahmadi (2002) introduced new priors for a covariance matrix and Bayesian hierarchical models for shrinking a covariance matrix towards structure. The (T,V)-reparameterization of Σ in (2.6) along with $\mathrm{P}(\Sigma) = \mathrm{P}(T|V)\mathrm{P}(V)$ provided a convenient framework for developing conditionally conjugate prior distributions for covariance matrices.

2.2 Regularized Estimation in the Presence of High Dimensionality

In recent years, data sets with high dimension and small sample size relative to dimension have become very common. Examples include gene expression arrays, spectroscopic imaging, numerical weather forecasting, and many others. Depending on the applications, the sparsity of the covariance matrix or the precision matrix Σ^{-1} is frequently imposed. Estimating large covariance matrices, where the dimension of the data p is comparable to or larger than the sample size n, has gained particular attention recently, since high-dimensional data are so common in applications. The regularization procedures have been generally used for the estimation of sparse covariance matrix or precision matrix.

Two broad groups of covariance estimations have emerged: those that rely on a natural ordering among variables, assuming that variables far apart in the ordering are only weakly correlated and those invariant under variable permutations, which will be discussed in more details below.

2.2.1 Approaches Rely on Natural Ordering among Variables

There are many applications that depend on the natural ordering among variables, which include regularizing the covariance matrix by banding or tapering (see, for example, Bickel & Levina (2004), Furrer & Bengtsson (2007), Bickel & Levina (2008b)) and regularizing Cholesky factor of the precision matrix. The sparsity in the inverse is usually introduced via the modified Cholesky decomposition (Pourahmadi (1999)). Bickel & Levina (2008b) considered estimating a covariance matrix of p variables with n observations by banding or tapering the sample covariance matrix. That is, banding the sample covariance matrix by

$$\hat{\Sigma}_{k,p} \equiv \hat{\Sigma}_k = B_k(\hat{\Sigma}_p),\ 0 \leq k < p, \tag{2.9}$$

or replacing $\hat{\Sigma}_p$ with $\hat{\Sigma}_p * R$, where $*$ denotes Schur (coordinate-wise) matrix multiplication and R is positive definite and symmetric. Estimating a banded version of precision matrix has also been proposed. They showed that banding the Cholesky factor produces a consistent estimator at various rates in the operator normal as long as $(\log p)^2/n \to 0$,

which implies that maximal and minimal eigenvalues of the estimates and Σ_p are close.

The methods of regularizing Cholesky factor of the precision matrix use the fact that the entries of the Cholesky factor have a regression interpretation. Therefore, the application of regularization tools, for example, the lasso and ridge penalties (Huang et al. (2006)), or the nested lasso penalty (Levina et al. (2008)) which mainly designed for the ordered variables situation, can be allowed. Wu & Pourahmadi (2003) proposed a k-diagonal banded estimator, which was obtained by local polynomial smoothing along the first k sub-diagonals of T and setting the rest to 0. An AIC or BIC penalty was suggested to select the number of k. They showed that the resulting estimate of the inverse was also k-banded and in addition, was element-wise consistent. Huang et al. (2006) proposed adding an L-1 penalty on the elements of the Cholesky factor T to the normal likelihood, which leads to Lasso-type shrinkage of the coefficients in T , and introduces zeros in T which can be placed in arbitrary locations, which is an advantage over Wu & Pourahmadi (2003). This approach is more flexible than banding the Cholesky factor, but the resulting estimate of the inverse may not have any zeros at all, hence, the sparsity is lost. No consistency results are available for this method. Also relying on the Cholesky decomposition and penalty function, Levina et al. (2008) imposed a banded structure on the Cholesky factor T, and selected the bandwidth adaptively for each row of T, by introducing a novel nested Lasso penalty on the coefficients of regressions that form the matrix T. It was shown that the structure are more flexibility than regular banding, but, unlike regular Lasso applied to the entries of the Cholesky factor, results in a sparse estimator for the precision matrix. Those above-mentioned methods are appropriate for a number of applications with ordered data, for example, climate data, spectroscopy data, time series data, etc. However, there are many applications, for example, gene expression arrays, of which the variables have no distance measure. Methods proposed to address those data will be discussed below.

2.2.2 Approaches Invariant to Variable Permutations

There is also an urge to construct estimators invariant under variable permutations. Several recent papers develop a sparse permutation-invariant estimate of the precision matrix. See for example, d'Aspremont et al. (2007), Rothman et al. (2008), Yuan & Lin (2007). The common approach of their methods is to add an L_1 (lasso) penalty on the

entries of the precision matrix to the normal likelihood, resulting in the shrinkage of some components to zero. In Rothman et al. (2008), it has been shown that the rate of convergence is driven by $(\log p)/n$. However, it is nontrivial to compute the estimator for high dimensions.

Sparsity in the inverse is particularly helpful in graphical models, since zeros in the inverse imply a graph structure. The parameter estimation and model selection in graphical model is equivalent to estimating parameters and identifying zeros in the precision matrix. Meinshausen & Bühlmann (2006) proposed a computationally attractive method for covariance selection that can be used for sparse high-dimensional graphs. They performed neighborhood selection with the LASSO for each node in the graph and combine the results to learn the structure of a Gaussian concentration graph model. They showed that the proposed neighborhood selection method is consistent for sparse high-dimensional graphs. However, the obstacle of the neighborhood selection method in Meinshausen & Bühlmann (2006) is that the model selection and parameter estimation are done separately. The parameters in the concentration matrix are typically estimated based on the model selected. Yuan & Lin (2007) proposed a penalized-likelihood method that does model selection and parameter estimation simultaneously in the Gaussian graphical model. They employed an L_1 penalty on the off-diagonal elements of the precision matrix, which is similar to the idea of the lasso in linear regression (Tibshirani (1996)). The lasso-type estimator $\hat{C}$ minimizes

$$-\log|C| + \mathrm{tr}(C\bar{A}) + \lambda \sum_{i \neq j} |c_{ij}|, \tag{2.10}$$

where $C = \Sigma^{-1}$ and the nonnegative garrote-type estimator minimizes

$$-\log|C| + \mathrm{tr}(C\bar{A}) + \lambda \sum_{i \neq j} \frac{c_{ij}}{\tilde{c}_{ij}}, \text{ subject to } c_{ij}/\tilde{c}_{ij} \geq 0. \tag{2.11}$$

Their methods leads to a sparse and shrinkage estimator of the concentration matrix that is positive definite, and thus conduct model selection and estimation simultaneously. In addition, they showed that although the implementation of the methods is nontrivial due to the positive constraint, the computation can be done effectively by taking advantage of the efficient maxdet algorithm developed in convex optimization.

Zou et al. (2006) applied L_1 penalty to loadings in the context of PCA to achieve

sparse representation. They proposed a method called sparse principal component analysis (SPCA) using the LASSO (elastic net) to produce modified principal components with sparse loadings. They showed that PCA can be formulated as a regression-type optimization problem, then sparse loadings can be obtained by imposing the LASSO (elastic net) constraint on the regression coefficients. They proposed efficient algorithms to realize SPCA for both regular multivariate data and gene expression arrays, and showed that the methods enjoyed advantages in aspects including computational efficiency, high explained variance and ability of identifying important variables.

Thresholding the sample covariance matrix in high-dimensional setting was thoroughly studied by Karoui (2008), Bickel & Levina (2008a), and Cai et al. (2008) with remarkable results for high-dimensional applications. One of the biggest advantages is its simplicity - hard thresholding carries no computation burden, unlike many other methods of covariance regularization. In the setting of "large sample size n, large dimensionality p", Karoui (2008) developed an estimator for sparse matrix by hard thresholding small entries of the sample covariance matrix and putting them to zero. The estimator was shown to be consistent in operator norm, under the condition that certain moments exist. Similarly, Bickel & Levina (2008a) proposed thresholding of the sample covariance matrix as a simple and permutation - invariant method of covariance regularization. The difference is that they developed a natural permutation-invariant notion of sparsity, which is more specialized then Karoui (2008)'s. It was shown that the thresholded estimate is consistent in the operator norm as long as the true covariance matrix is sparse, the variables are Gaussian or sub-Gaussian, and $(\log p)/n \to 0$, and obtain explicit rates. A potential disadvantage is the loss of positive definiteness. However, the proposed estimator for a suitably sparse class of matrices is shown to be consistent under certain conditions, the estimator will be positive definite with probability tending to 1. Despite the consistency of estimators in operator norm by Karoui (2008) and Bickel & Levina (2008a), and explicitness of the rates of convergence, it is not clear whether any of these rates of convergence are optimal. Cai et al. (2008) established the optimal rate of convergence for estimating the covariance matrix and the precision matrix over a wide range of classes of covariance matrices. Both the operator norm and Frobenius norm were considered. However, all the aforementioned methods are not directly applicable to estimating sparse precision matrix when the dimensionality p_n is greater than the sample size n.

2.3 Nonparametric Models for the Covariance Matrix

In order to balance between variability and bias of the covariance estimators, it is reasonable to contract attention to covariance structures suggested by the data. To this end, nonparametric approaches for estimating the covariance structures are useful not only as a guide to the formulation of parametric models but also as the basis for formal inference without requiring additional parametric assumptions. Glasbey (1988), Shapiro & Botha (1991), Sampson & Guttorp (1992), Hall et al. (1994), and Hall & Patil (1994) proposed several nonparmetric models to estimate the covariance matrices. However, most nonparametric estimators of covariance matrices were developed either for stationary processes or without heeding the positive-definiteness constraint. Diggle & Verbyla (1998) introduced a nonparametric estimator for the covariance structure of longitudinal data without assuming stationarity. However, their estimator, based on kernel weighted local linear regression smoothing of sample variogram ordinates and of squared residuals, is not guaranteed to be positive definite.

In the analysis of longitudinal data, it is an important issue to estimate the covariance functions. It features notably in forecasting the trajectory of an individual response over time and is closely related with improving the efficiency of regression coefficients estimated. Nevertheless, due to the fact that longitudinal data are frequently collected at irregular and possibly subject-specific time points, great challenges would be arised in estimating the covariance function. Interest in this kind of challenges has surged in the recent literature. Following Fan & Zhang (2000)'s two-step estimation of functional linear models, Wu & Pourahmadi (2003) proposed nonparametric estimators of the covariance matrices which are guaranteed to be positive definite. Specifically, they applied nonparametric smoothing, by local polynomials, to the first few subdiagonals of the Cholesky factor and set to zero the remaining subdiagonals, thereby restricting T to be a banded lower triangular matrix. The asymptotic results for the local polynomial estimators of components of the covariance matrix were established. They showed that the single matrix element estimates converge to their population values in probability, with $p_n \to \infty$ at a certain rate determined by the spline smoothers used. The limitation of their method is that the proposed approach can deal with only balanced or nearly balanced longitudinal data.

There are few references available for nonparametric models for a covariance func-

tion. A class of semiparametric models for the covariance function was proposed by Fan et al. (2007). They considered a semiparametric varying-coefficient partially linear model:

$$y(t) = \mathbf{x}(t)^T \alpha(t) + \mathbf{z}(t)^T \beta + \varepsilon(t), \tag{2.12}$$

where $\alpha(t)$ consists of p unknown smooth functions, β is a q-dimensional unknown parameter vector, and $\mathrm{E}\{\varepsilon(t)|\mathrm{x}(t),\mathrm{z}(t)\} = 0$. Nonparametric models for longitudinal data (Lin & Carroll (2001); Wang (2003)) can be viewed as a special case of model (2.12). Focusing on parsimonious modeling of the covariance function of the random error process $\varepsilon(t)$ for the analysis of longitudinal data, when observations are collected at irregular and possibly subject-specific time points, they imposed a parametric correlation structure (i.e. $\mathrm{corr}\{\varepsilon(s),\varepsilon(t) = \rho(t,s,\theta)\}$, where $\rho(t,s,\theta)$ is a positive definite function of s and t) while allowing a nonparametric variance function (i.e. $\mathrm{Var}\{\varepsilon(t)|\mathrm{x}(t),\mathrm{z}(t)\} = \sigma^2(t)$). The semiparametric model guarantees positive definiteness for the resulting estimate; it retains the flexibility of nonparametric modeling and parsimony and ease of interpretation of parametric modeling. A kernel estimator was developed for the estimation of the nonparametric variance function $\sigma^2(t)$. They further developed estimation procedures for parameters θ in correlation structure using quasi-likelihood and minimum generalized variance approaches. They introduced a semiparametric varying coefficient partially linear model for longitudinal data and proposed an estimation procedure for model coefficients $\alpha(t)$ and β by using a profile weighted least squares approach. Sampling properties of the proposed estimation procedures were studied and asymptotic normality of the resulting estimators was established.

Yin et al. (2010) proposed a nonparametric model for the conditional covariance matrix, which can be regarded as a natural extension of existing nonparametric models for conditional variance. For two random variables X and U, they modeled the conditional covariance of X given U as $\mathrm{Cov}(X|U) = \Sigma(U)$, where the component was assumed to be an unknown but smooth function of U. They developed a Nadaraya-Watson (NW) kernel estimators for both the conditional mean $m(u)$ and the conditional covariance $\Sigma(u)$. Specifically, the kernel method was to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\big\{X_i - m(u)\big\}^T \Sigma^{-1}(u)\big\{X_i - m(u)\big\} - \log\big(|\Sigma^{-1}(u)|\big)\Big] K_h(U_i - u). \tag{2.13}$$

The resulting NW kernel estimator was obtained, of which the asymptotic bias, variance and in addition asymptotic normality were also derived. They found that without knowing the true regression function, the conditional covariance matrix can be asymptotically estimated as well as if the true regression function was known in advance.

Chapter 3

Nonparametric Covariance Regression Models

This chapter presents a general approach to estimate the conditional covariance matrix in the context of nonparametric regression models. Two covariance regression models are considered: one with continuous explanatory variable only and the other with both continuous and discrete explanatory variables. We propose an estimation procedure for each model via random-effects representation and an EM-algorithm. Monte Carlo simulation studies are conducted to assess the finite sample performance of the proposed procedure and a real data example is used to illustrate the proposed methodology.

3.1 Covariance Regression Model I

Let $\mathbf{y} \in \mathbb{R}^p$ be a random multivariate response vector and $\mathbf{x} \in \mathbb{R}^q$ be a vector of predict variables. The goal is to provide a model and the method to estimate $\mathrm{Cov}(\mathbf{y}|\mathbf{x}) = \Sigma_{\mathbf{x}}$, the conditional covariance matrix of $\mathbf{y}$ given $\mathbf{x}$. Now considering a cubic spline $\mathbf{S}(\mathbf{x})$ with knots $\kappa_1, \ldots, \kappa_J$, which has linear, quadratic and cubic terms on $\mathbf{x}$, and one term of the form $(\mathbf{x} - \kappa_j)_+^3$ for each knot, the proposed covariance regression model is given by

$$\Sigma_{\mathbf{x}} = \mathbf{A} + \mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{B}^T, \tag{3.1}$$

where $\mathbf{S}(\mathbf{x}) = [\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, (\mathbf{x} - \kappa_1)_+^3, \cdots, (\mathbf{x} - \kappa_J)_+^3]$ is a $q \times (J+3)$ matrix, $\mathbf{A}$ is a $p \times p$ positive-definite matrix, and $\mathbf{B}$ is a $p \times q$ matrix. The resulting covariance function is

positive definite for all $\mathbf{x}$ and represents the conditional covariance as a "baseline" covariance matrix $\mathbf{A}$ plus a $p \times p$ nonnegative definite matrix depending on $\mathbf{x}$. The $p \times q$ parameters of $\mathbf{B}$ have an explicit explanation of the heteroscedasticity among the p variables of $\mathbf{y}$. In addition, the model has a random-effect representation, allowing for the maximum likelihood parameter estimation via the EM-algorithm.

3.1.1 Model Interpretation

Consider a model for data $\mathbf{Y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ observed under the conditions $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T$ with $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})^T, \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^T$. The model has the following form:

$$\mathbf{y}_i \quad = \quad \boldsymbol{\mu}_{\mathbf{x}_i} + \mathbf{g}_i(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \tag{3.2}$$

where $\boldsymbol{\mu}_{\mathbf{x}_i}$ is a constant vector, $\boldsymbol{\varepsilon}_i$ is a random vector, and $\mathbf{g}_i(\mathbf{x}_i)$ is a vector of components $g_{ij}(\mathbf{x}_i)$, $j = 1, 2, \ldots, p$. It is assumed that each $g_{ij}(\mathbf{x}_i)$ is a nonparametric function and can be expressed as a linear combination of a set of splines basis in terms of the unit's explanatory vector $\mathbf{x}_i$. Let,

$$\mathbf{s}(\mathbf{x}_i) = \gamma_{i1}\mathbf{x}_i + \gamma_{i2}\mathbf{x}_i^2 + \gamma_{i3}\mathbf{x}_i^3 + \sum_{j=1}^{J} \gamma_{i,3+j}(\mathbf{x}_i - \kappa_j)_+^3 = \mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i \ , \tag{3.3}$$

with a random vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{3+J})^T$. Then $g_{ij}(\mathbf{x}_i)$ can be expressed in terms of $\mathbf{s}(\mathbf{x}_i)$. That is,

$$g_{ij}(\mathbf{x}_i) = \mathbf{b}_j^T \mathbf{s}(\mathbf{x}_i) = \mathbf{b}_j^T \mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i, \tag{3.4}$$

where $\mathbf{b}_j^T$ is the j-th row of a $p \times q$ matrix $\mathbf{B}$, and $\mathbf{g}_i(\mathbf{x}_i) = \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i$, an explanatory vector for unit i.

Therefore, $\mathbf{y}_i$ can be expressed as

$$\mathbf{y}_i = \boldsymbol{\mu}_{\mathbf{x}_i} + \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i. \tag{3.5}$$

It is assumed that the random variables satisfy

$$\begin{aligned} \mathrm{E}(\boldsymbol{\varepsilon}_i) &= \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \mathbf{A}; && (3.6)\\ \mathrm{E}(\boldsymbol{\gamma}_i) &= \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\gamma}_i) = \mathbf{I},\ \mathrm{E}(\boldsymbol{\gamma}_i\boldsymbol{\varepsilon}_i^T) = \mathbf{0}, && (3.7)\\ i &= 1,2,\ldots,n. \end{aligned}$$

Let $\{\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_p\}$ be the rows of $\mathbf{B}$. The covariance regression model gives

$$\begin{aligned} \mathrm{Var}(y_{ij}|\mathbf{x}_i) &= a_{j,j}+\mathbf{b}_j^T\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{b}_j,\\ \mathrm{Cov}(y_{ij},y_{ik}|\mathbf{x}_i) &= a_{j,k}+\mathbf{b}_j^T\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{b}_k. \end{aligned}$$

The above parameterizations of the variance indicates that the variance in each element of response vector $\mathbf{y}$ to be increasing in the elements of explanatory vector $\mathbf{x}$, if all elements of $\mathbf{b}_j$ are of the same signs. The minimum variance is obtained when $\mathbf{x}=\mathbf{0}$.

Additionally, the model given by equation (3.5) can be regarded as a factor analysis model and thus has a random-effects representation. The $p\times q$ parameters of matrix $\mathbf{B}$ have a direct interpretation in terms of how heteroscedasticity co-occurs among the p variables of $\mathbf{y}$. In order to see how the random-effects representation affects the variance, let $\{\mathbf{b}_1,\ldots,\mathbf{b}_p\}$ be the rows of $\mathbf{B}$. The covariance regression model (3.5) can then be expressed as

$$\begin{pmatrix} y_{i1}-\mu_{x_i1}\\ \vdots\\ y_{ip}-\mu_{x_ip}\end{pmatrix} = \begin{pmatrix} \mathbf{b}_1^T\mathbf{S}(\mathbf{x}_i)\\ \vdots\\ \mathbf{b}_p^T\mathbf{S}(\mathbf{x}_i)\end{pmatrix}\boldsymbol{\gamma}_i + \begin{pmatrix}\varepsilon_{i1}\\ \vdots\\ \varepsilon_{ip}\end{pmatrix}. \qquad (3.8)$$

where $\boldsymbol{\gamma}_i$ expresses additional variability beyond that represented by $\boldsymbol{\varepsilon}_i$. The vectors $\{\mathbf{b}_1,\ldots,\mathbf{b}_p\}$ describe how this additional variability is displayed across the p different response variables. Small values of $\mathbf{b}_j$ indicate little heteroscedasticity in y_j as a function of $\mathbf{x}$, while big values of $\mathbf{b}_j$ indicate large heteroscedasticity in $\mathbf{y}_j$. $\mathbf{y}_j$ and $\mathbf{y}_k$ become more positively or more negatively correlated, respectively, as their variances increase, when vectors $\mathbf{b}_j$ and $\mathbf{b}_k$ begin either in the same or opposite direction.

The random effects representation of the model gives that the resulting covariance

matrix for $\mathbf{y}_i$ given $\mathbf{x}_i$ equals to

$$\begin{aligned}
\text{Cov}(\mathbf{y}_i|\mathbf{x}_i) &= \text{E}\left\{(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i})(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i})^T\right\} \\
&= \text{E}\left\{\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T+\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\boldsymbol{\varepsilon}_i^T+\boldsymbol{\varepsilon}_i\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T+\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T\right\} \\
&= \mathbf{BS}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}+\mathbf{A} \\
&= \Sigma_{\mathbf{x}_i}, \qquad (3.9)
\end{aligned}$$

which is exactly the same form as that proposed in (3.1).

3.1.2 Parameter Estimation with the EM-algorithm

Suppose that $\{\mathbf{y}_i,\mathbf{x}_i,i=1,2,\ldots,n\}$ is a random sample from the population $\{\mathbf{y},\mathbf{x}\}$, where $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{ip})^T$ and $\mathbf{x}_i=(x_{i1},x_{i2},\ldots,x_{iq})^T$. We now consider parameter estimation based on observed data $\mathbf{Y}=(\mathbf{y}_1^T,\ldots,\mathbf{y}_n^T)^T$ given $\mathbf{X}=(\mathbf{x_1^T},\ldots,\mathbf{x_n^T})^T$. We assume normal models for all error terms:

$$\begin{aligned}
\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n &\sim N_d(\mathbf{0},\mathbf{I}_{d\times d}), \\
\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n &\sim N_p(\mathbf{0},\mathbf{A}_{p\times p}), \qquad (3.10)
\end{aligned}$$

for the proposed model (3.5), where $\mathbf{S}(\mathbf{x}_i)=\left(\mathbf{x}_i,\mathbf{x}_i^2,\mathbf{x}_i^3,\ (\mathbf{x}_i-\kappa_1)_+^3,\ \cdots,\ (\mathbf{x}_i-\kappa_J)_+^3\right)$, a $q\times d$ matrix with $d=3+J$, $\mathbf{y}_i,\boldsymbol{\mu}_{\mathbf{x}_i}$ and $\boldsymbol{\varepsilon}_i$ are $p\times 1$ vectors, $\boldsymbol{\gamma}_i$ is a $d\times 1$ vector, and $\mathbf{B}$ is $p\times q$ matrix.

Assume $\{\boldsymbol{\mu}_{\mathbf{x}},\mathbf{x}\in\mathscr{X}\}$ are known and let $\mathbf{E}=(\mathbf{e}_1^T,\cdots,\mathbf{e}_n^T)^T$ be the matrix of residuals $\mathbf{e_i}=\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i}$, $i=1,2,\ldots,n$. The log-likelihood of the parameters based on $\mathbf{X}$ and $\mathbf{E}$ is:

$$\begin{aligned}
&\ell(\mathbf{A},\mathbf{B}:\mathbf{E},\mathbf{X}) \qquad (3.11)\\
&= c-\frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{A}+\mathbf{BS}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T|-\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left[\left\{\mathbf{A}+\mathbf{BS}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\right\}^{-1}\mathbf{e}_i\mathbf{e}_i^T\right].
\end{aligned}$$

It can be shown that the maximum likelihood estimates of $\mathbf{A}$ and $\mathbf{B}$ satisfy the following equations:

$$\sum_i\hat{\Sigma}_{\mathbf{x}_i}^{-1} = \sum_i\hat{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{e}_i\mathbf{e}_i^T\hat{\Sigma}_{\mathbf{x}_i}^{-1},$$

$$\sum_i \hat{\Sigma}_{\mathbf{x}_i}^{-1}\hat{\mathbf{B}}\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T = \sum_i \hat{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{e}_i\mathbf{e}_i^T\hat{\Sigma}_{\mathbf{x}_i}^{-1}\hat{\mathbf{B}}\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T, \tag{3.12}$$

where $\hat{\Sigma}_{\mathbf{x}_i} = \hat{\mathbf{A}} + \hat{\mathbf{B}}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\hat{\mathbf{B}}^T$. We may use kronecker product to get the close-form expressions for $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. However, the direct maximization of (3.11) from kronecker product is complicated.

The maximization likelihood estimation via simple iterative methods. For example, the EM-algorithm, is straightforward. This method relies on the conditional distribution of $\{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n\}$ given $\{\mathbf{A},\mathbf{B},\mathbf{Y},\mathbf{X}\}$, which can be derived as follows.

The conditional distribution of $\mathbf{y}_i$ given $(\boldsymbol{\gamma}_i,\mathbf{x}_i,\mathbf{A},\mathbf{B})$ and the marginal distribution of $\boldsymbol{\gamma}_i$ can be expressed as

$$\begin{aligned}\mathbf{y}_i \mid \boldsymbol{\gamma}_i,\mathbf{x}_i,\mathbf{A},\mathbf{B} &\sim N_p(\boldsymbol{\mu}_{\mathbf{x}_i}+\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i,\mathbf{A}),\\ \boldsymbol{\gamma}_i &\sim N_d(\mathbf{0},\mathbf{I}_{d\times d}),\end{aligned}$$

then the conditional distribution of $(\mathbf{y}_i,\boldsymbol{\gamma}_i)$ given $(\mathbf{x}_i,\mathbf{A},\mathbf{B})$ can be expressed as

$$\begin{aligned}&f(\mathbf{y}_i,\boldsymbol{\gamma}_i \mid \mathbf{x}_i,\mathbf{A},\mathbf{B})\\ &= \exp\Big\{-\frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i}-\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i)^T\mathbf{A}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i}-\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i)-\frac{1}{2}\boldsymbol{\gamma}_i^T\boldsymbol{\gamma}_i\Big\}\\ &\quad\times(2\pi)^{-\frac{p+d}{2}}|\mathbf{A}|^{-\frac{1}{2}}.\end{aligned}$$

Thus the conditional distribution of $\mathbf{y}_i$ given $(\mathbf{x}_i,\mathbf{A},\mathbf{B})$ is known as

$$\begin{aligned}&f(\mathbf{y}_i \mid \mathbf{x}_i,\mathbf{A},\mathbf{B})\\ &= \int f(\mathbf{y}_i,\boldsymbol{\gamma}_i \mid \mathbf{x}_i,\mathbf{A},\mathbf{B})d\boldsymbol{\gamma}_i\\ &= \int (2\pi)^{-\frac{p+d}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\Big[-\frac{1}{2}\big\{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{e}_i-\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\\ &\quad(\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i))^{-1}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i\big\}\Big]\\ &\quad\times\exp\Big[-\frac{1}{2}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_{\boldsymbol{\gamma}_i})^T(\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i))(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_{\boldsymbol{\gamma}_i})\Big]d\boldsymbol{\gamma}_i\\ &= (2\pi)^{-\frac{p}{2}}|\mathbf{A}|^{-\frac{1}{2}}|\Sigma_{\boldsymbol{\gamma}_i}|^{\frac{1}{2}}\exp\Big[-\frac{1}{2}\big\{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{e}_i-\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\\ &\quad(\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i))^{-1}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i\big\}\Big]\end{aligned}$$

$$= (2\pi)^{-\frac{p}{2}}|\mathbf{A}|^{-\frac{1}{2}}|\Sigma_{\boldsymbol{\gamma}_i}|^{\frac{1}{2}}\exp\Big[-\frac{1}{2}\big\{\mathbf{e}_i^T\mathbf{A}^{-1}(\mathbf{I}-\mathbf{B}\mathbf{S}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}^{-1}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1})\mathbf{e}_i\big\}\Big],$$

where

$$\begin{aligned}\Sigma_{\boldsymbol{\gamma}_i} &= \big\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\big\}^{-1},\\ \boldsymbol{\mu}_{\boldsymbol{\gamma}_i} &= \big\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\big\}^{-1}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i.\end{aligned}$$

Hence the conditional distribution of $\boldsymbol{\gamma}_i$ given $\{\mathbf{Y},\mathbf{X},\mathbf{A},\mathbf{B}\}$ is expressed as

$$\begin{aligned}&f(\boldsymbol{\gamma}_i \mid \mathbf{Y},\mathbf{X},\mathbf{A},\mathbf{B})\\ &= \frac{f(\mathbf{y}_i,\boldsymbol{\gamma}_i \mid \mathbf{x}_i,\mathbf{A},\mathbf{B})}{f(\mathbf{y}_i \mid \mathbf{x}_i,\mathbf{A},\mathbf{B})}\\ &= (2\pi)^{-\frac{d}{2}}|\Sigma_{\boldsymbol{\gamma}_i}|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\big[\boldsymbol{\gamma}_i^T\left\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\right\}\boldsymbol{\gamma}_i\\ &\quad -\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i+\boldsymbol{\mu}_{\mathbf{x}_i}^T\Sigma_{\boldsymbol{\gamma}_i}\boldsymbol{\mu}_{\mathbf{x}_i}\big]\Big).\end{aligned}$$

Therefore, the conditional distribution of $\{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n\}$ given $\{\mathbf{A},\mathbf{B},\mathbf{Y},\mathbf{X}\}$ is as follows:

$$\begin{aligned}\big(\boldsymbol{\gamma}_\mathbf{i}|\mathbf{Y},\mathbf{X},\mathbf{A},\mathbf{B}\big) &\sim N_d(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\Sigma_{\boldsymbol{\gamma}_i}),\text{ where}\\ \Sigma_{\boldsymbol{\gamma}_i} &= \big\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\big\}^{-1},\\ \boldsymbol{\mu}_{\boldsymbol{\gamma}_i} &= \Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i. \qquad (3.13)\end{aligned}$$

The EM-algorithm alternates between computing the expectation of the complete data log-likelihood evaluated using the current estimate for the latent variables and computing parameters maximizing the expected log-likelihood. The complete data log-likelihood $\ell(\mathbf{A},\mathbf{B})$ is given as

$$\begin{aligned}&\ell(\mathbf{A},\mathbf{B})\\ &= -\frac{1}{2}\left[np\log(2\pi)+n\log|\mathbf{A}|+\sum_{i=1}^{n}\big\{\mathbf{e}_i-\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i\big\}^T\mathbf{A}^{-1}\big\{\mathbf{e}_i-\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i\big\}\right]. \qquad (3.14)\end{aligned}$$

Given current estimates $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ of $(\mathbf{A},\mathbf{B})$, the classical EM approach is to maximize

the expected data log-likelihood $\mathrm{E}\{\ell(\mathbf{A},\mathbf{B})|\hat{\mathbf{A}},\hat{\mathbf{B}}\}$,

$$\begin{aligned}&\mathrm{E}\{\ell(\mathbf{A},\mathbf{B})|\hat{\mathbf{A}},\hat{\mathbf{B}}\}\\&=-\frac{1}{2}\left(np\log(2\pi)+n\log|\mathbf{A}|+\sum_{i=1}^{n}\mathrm{E}\left[\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}|\hat{\mathbf{A}},\hat{\mathbf{B}}\right]\right).\end{aligned}$$

To drive the EM formula, the representation of $\mathrm{E}\{\ell(\mathbf{A},\mathbf{B})|\hat{\mathbf{A}},\hat{\mathbf{B}}\}$ needs to be simplified.

Denote $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}=\mathrm{E}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i)$ and $\Sigma_{\boldsymbol{\gamma}_i}=\mathrm{Var}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i)$ as the conditioned mean and variance of $\boldsymbol{\gamma}_i$ given $\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i$, which will be used in the following analysis. It is easy to verify that

$$\begin{aligned}&\mathrm{E}\left[\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}|\hat{\mathbf{A}},\hat{\mathbf{B}}\right]\\&=\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}+\mathrm{E}\left[\left\{\mathbf{BS}(\mathbf{x}_i)\tilde{\boldsymbol{\gamma}}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{BS}(\mathbf{x}_i)\tilde{\boldsymbol{\gamma}}_i\right\}\right]\\&=\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}+\mathrm{tr}\left\{\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{BS}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\right\}.\end{aligned}$$

where $\tilde{\boldsymbol{\gamma}}_i=\boldsymbol{\gamma}_i-\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$. By the factorization $\Sigma_{\boldsymbol{\gamma}_i}=\mathbf{K}_i\mathbf{K}_i^T$ of $\Sigma_{\boldsymbol{\gamma}_i}$ where $\mathbf{K}_i$ is a $d\times d$ matrix, the above second term can be represented as

$$\mathrm{tr}\left\{\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{BS}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\right\}=\mathrm{tr}\left[\left\{\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right\}^T\mathbf{A}^{-1}\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right].$$

Thus,

$$\begin{aligned}&\mathrm{E}\left[\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}|\hat{\mathbf{A}},\hat{\mathbf{B}}\right]\\&=\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}\right\}+\mathrm{tr}\left[\left\{\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right\}^T\mathbf{A}^{-1}\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right]\\&=\mathrm{tr}\left[\left\{(\mathbf{e}_i,\mathbf{0})-\mathbf{BS}(\mathbf{x}_i)(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\mathbf{K}_i)\right\}^TA^{-1}\left\{(\mathbf{e}_i,\mathbf{0})-\mathbf{BS}(\mathbf{x}_i)(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\mathbf{K}_i)\right\}\right]\\&=\mathrm{tr}\left\{(\tilde{\mathbf{E}}_i-\mathbf{B}\tilde{\mathbf{X}}_i)^TA^{-1}(\tilde{\mathbf{E}}_i-\mathbf{B}\tilde{\mathbf{X}}_i)\right\},\end{aligned}$$

where $\tilde{\mathbf{X}}_i=\mathbf{S}(\mathbf{x}_i)(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\mathbf{K}_i)$ and $\tilde{\mathbf{E}}_i=(\mathbf{e}_i,\mathbf{0})$ has only a nonzero column. We also have the sum in a form of a matrix trace,

$$\sum_{i=1}^{n}\mathrm{E}\left\{(\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i)^T\mathbf{A}^{-1}(\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i)|\hat{\mathbf{A}},\hat{\mathbf{B}}\right\}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \mathrm{tr}\big\{(\tilde{\mathbf{E}}_i - \mathbf{B}\tilde{\mathbf{X}}_i)^T A^{-1}(\tilde{\mathbf{E}}_i - \mathbf{B}\tilde{\mathbf{X}}_i)\big\} \\
&= \mathrm{tr}\big\{(\tilde{\mathbf{E}} - \mathbf{B}\tilde{\mathbf{X}})^T A^{-1}(\tilde{\mathbf{E}} - \mathbf{B}\tilde{\mathbf{X}})\big\},
\end{aligned}
$$

where

$$
\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n), \quad \tilde{\mathbf{E}} = (\tilde{\mathbf{E}}_1, \ldots, \tilde{\mathbf{E}}_n).
$$

The expected value of the complete data log-likelihood can be simply written as

$$
\mathrm{E}\left\{\ell(\mathbf{A},\mathbf{B})|\hat{\mathbf{A}},\hat{\mathbf{B}}\right\} = -\frac{1}{2}\left[np\log(2\pi) + n\log|\mathbf{A}| + \mathrm{tr}\big\{(\tilde{\mathbf{E}} - \mathbf{B}\tilde{\mathbf{X}})^T \mathbf{A}^{-1}(\tilde{\mathbf{E}} - \mathbf{B}\tilde{\mathbf{X}})\big\}\right],
$$

which is the likelihood for multivariate normal regression. Clearly, the maximizer of the $\mathrm{E}\{\ell(\mathbf{A},\mathbf{B})|\hat{\mathbf{A}},\hat{\mathbf{B}}\}$ is given by

$$
\begin{aligned}
\tilde{\mathbf{B}} &= \tilde{\mathbf{E}}\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1}, && (3.15)\\
\tilde{\mathbf{A}} &= (\tilde{\mathbf{E}} - \hat{\mathbf{B}}\tilde{\mathbf{X}})(\tilde{\mathbf{E}} - \hat{\mathbf{B}}\tilde{\mathbf{X}})^T/n. && (3.16)
\end{aligned}
$$

The above EM formula can be further simplified. Practically, it is not required to factorize the matrix $\Sigma_{\boldsymbol{\gamma}_i}$, which will be shown below. For simplicity, let $\mathbf{u}_i = \mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$. By definition,

$$
\begin{aligned}
\tilde{\mathbf{E}}\tilde{\mathbf{X}}^T &= \sum_{i=1}^{n} \tilde{\mathbf{E}}_i\tilde{\mathbf{X}}_i^T = \sum_{i=1}^{n} \mathbf{e}_i\mathbf{u}_i^T = \mathbf{E}\mathbf{U}^T, \\
\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T &= \sum_{i=1}^{n} \left\{\mathbf{u}_i\mathbf{u}_i^T + \mathbf{S}(\mathbf{x}_i)\mathbf{K}_i\mathbf{K}_i^T\mathbf{S}(\mathbf{x}_i)^T\right\} = \mathbf{U}\mathbf{U}^T + \sum_{i=1}^{n} \mathbf{S}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
(\tilde{\mathbf{E}} - \hat{\mathbf{B}}\tilde{\mathbf{X}})(\tilde{\mathbf{E}} - \hat{\mathbf{B}}\tilde{\mathbf{X}})^T &= \sum_{i=1}^{n} (\mathbf{e}_i - \hat{\mathbf{B}}\mathbf{u}_i)(\mathbf{e}_i - \hat{\mathbf{B}}\mathbf{u}_i)^T + \hat{\mathbf{B}}\Big\{\sum_{i=1}^{n} S(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}S(\mathbf{x}_i)^T\Big\}\hat{\mathbf{B}}^T \\
&= (\mathbf{E} - \hat{\mathbf{B}}\mathbf{U})(\mathbf{E} - \hat{\mathbf{B}}\mathbf{U})^T + \hat{\mathbf{B}}\Big\{\sum_{i=1}^{n} \mathbf{S}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T\Big\}\hat{\mathbf{B}}^T.
\end{aligned}
$$

The details of the EM algorithm for the multivariate normal regression are described as follows. Let $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ be the current estimation.

EM algorithm for the nonparametric regression model (3.5).

1. Estimate $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} = \mathrm{E}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i,\mathbf{x}_i)$ and $\Sigma_{\boldsymbol{\gamma}_i} = \mathrm{Var}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i,\mathbf{x}_i)$.
2. Compute $\mathbf{u}_i = \mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$, $\mathbf{R} = \mathbf{E} - \hat{\mathbf{B}}\mathbf{U}$, and $\mathbf{M} = \sum_{i=1}^{n} \mathbf{S}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T$.
3. Update $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ by

$$\hat{\mathbf{A}} = \left(\mathbf{R}\mathbf{R}^T + \hat{\mathbf{B}}\mathbf{M}\hat{\mathbf{B}}^T\right)/n, \quad \hat{\mathbf{B}} = \mathbf{E}\mathbf{U}^T(\mathbf{U}\mathbf{U}^{\mathbf{T}} + \mathbf{M})^{-\mathbf{1}}.$$

This procedure is repeated until a desired convergence criterion has been met.

3.2 Theoretical Properties

The asymptotic properties of the maximum likelihood estimators (MLEs) have been extensively studied. In this section, the consistency and asymptotic normality properties of the MLEs for the proposed nonparametric covariance regression model will be discussed.

Let Θ be a subset of $\mathbb{R}^m$. Let $P_\theta : \theta \in \Theta$ be a family of distributions on $(\Omega_{\mathbf{z}}, \mathscr{F}_{\mathbf{z}})$. Let μ be a σ-finite measure on $(\Omega_{\mathbf{z}}, \mathscr{F}_{\mathbf{z}})$. Suppose $P_\theta \ll \mu, \forall \theta \in \Theta$ and denote by $f_\theta = dP_\theta/d\mu$ the derivative of P_θ with respect to μ. The maximum likelihood estimate of θ is defined by

$$\hat{\theta} = \arg\max \left\{ \sum_{i=1}^{n} \log f_\theta(\mathbf{z}_i), \theta \in \Theta \right\}.$$

Lemma 3.2.1 (Cramér's Consistency)**.** *Suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d. from $f_\theta(\mathbf{z}), \theta \in \Theta$. Let $S(\theta,\mathbf{z}) = \frac{\partial}{\partial \theta^T} \log f_\theta(\mathbf{z})$, which is the score function. Let $E_n\{S(\theta,\mathbf{z})\} = \frac{1}{n}\sum_{i=1}^{n} S(\theta,\mathbf{z}_i)$. Suppose θ_0 is the true parameter and furthermore,*

a. $S(\theta,\mathbf{z})$ is continuous in θ;

b. $E\{S(\theta_0,\mathbf{z})\} \equiv 0$;

c. $E\{S(\theta,\mathbf{z})\}$ is differentiable at $\theta = \theta_0$ and the derivative matrix is negative definite;

d. In the neighborhood of θ_0, $E_n\{S(\theta,\mathbf{z})\}$ *converges in probability uniformly to* $E\{S(\theta,\mathbf{z})\}$. *In other words,* $\exists$ *a neighborhood G of* θ_0, *such that*

$$\sup_{\theta\in G}\|E_n\{S(\theta,\mathbf{z})\}-E\{S(\theta,\mathbf{z})\}\|\to_P 0;$$

Then, there is a sequence $\{\hat{\theta}_n : n=1,2,\ldots\}$ *such that*

i. $P\big(\hat{\theta}_n$ *is a solution to* $E_n\{S(\theta,\mathbf{z})\}=0\big)\to 1$;

ii. $\hat{\theta}_n\to_P\theta_0$.

The details of Lemma 3.2.1 can be found in several references (see, for example, Cramér (1946) and Lehmann (1998)).

Considering the nonparametric regression model (3.5), $\mathbf{y}_i|\mathbf{x}_i$'s are independent and identically distributed from multivariate normal, with mean $\boldsymbol{\mu}_{\mathbf{x}_i}$ and covariance $\mathbf{A}+\mathbf{BS}(\mathbf{x}_i)\mathbf{S}^T(\mathbf{x}_i)\mathbf{B}^T$, $i=1,2\ldots,n$. It is very challenging in establishing sampling properties of the resulting estimator of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. To simplify the theoretical proofs, we impose the following assumptions:

Assumption A1. The collected data is a random sample from model (3.2), in which $\mathbf{g}_i(\mathbf{x}_i)$ can be represented as $\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i$ for a set of finite pre-specified basis functions. That is, model (3.5) is a correct model.

Assumption A2. Assume that the mean $\boldsymbol{\mu}_{\mathbf{x}_i}$ is known. Without loss of generality, it is assumed that $\boldsymbol{\mu}_{\mathbf{x}_i}=\mathbf{0}$.

Assumption A1 implies that $\mathbf{g}_i(\mathbf{x}_i)$ lies in the functional space spanned by a set of the pre-specified basis functions. Thus, we do not need to analyze the approximation error of $\mathbf{g}_i(\mathbf{x}_i)$ in the theoretical proofs. The assumption A2 allows us to ignore the estimation error due to $\hat{\boldsymbol{\mu}}_{\mathbf{x}_i}$, and further simplify the proof.

Under Assumptions A1 and A2, the log-likelihood can be written as:

$$\begin{aligned}&\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})\\ &= \quad c-\frac{1}{2}\log|\mathbf{A}+\mathbf{BS}(\mathbf{x})\mathbf{S}^T(\mathbf{x})\mathbf{B}^T|-\frac{1}{2}\mathrm{tr}\left[\left\{\mathbf{A}+\mathbf{BS}(\mathbf{x})\mathbf{S}^T(\mathbf{x})\mathbf{B}^T\right\}^{-1}\mathbf{y}\mathbf{y}^T\right].\end{aligned}\quad(3.17)$$

Based on Lemma 3.2.1, for model (3.5) with the parameters $\theta = (\mathbf{A}, \mathbf{B})$ and the random variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ can be proved to be consistent. The Theorem 3.2.1 shows the consistent property of the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

Theorem 3.2.1 (Consistency)**.** *Let* $\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y}) = \frac{\partial}{\partial(\mathbf{A},\mathbf{B})} \log f_{\mathbf{A},\mathbf{B}}(\mathbf{x}, \mathbf{y})$*, which is the score function. Let* $E_n\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}_i, \mathbf{y}_i)$*. Denote by* $(\mathbf{A}_0, \mathbf{B}_0)$ *the true parameters. Under Assumptions A1 and A2,* $(\hat{\mathbf{A}}, \hat{\mathbf{B}}) \to_P (\mathbf{A}_0, \mathbf{B}_0)$.

Proof. The proof consists of validation of Conditions (a)—(d) step by step. Specifically, we will check the following conditions one by one:

a. $\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})$ is continuous in $(\mathbf{A}, \mathbf{B})$;

b. $\mathrm{E}\{\mathbf{S}(\mathbf{A}_0, \mathbf{B}_0, \mathbf{x}, \mathbf{y})\} \equiv 0$;

c. $\mathrm{E}\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\}$ is differentiable at $(\mathbf{A}, \mathbf{B}) = (\mathbf{A}_0, \mathbf{B}_0)$ and the derivative matrix is negative definite;

d. In the neighborhood of $(\mathbf{A}_0, \mathbf{B}_0)$, $\mathrm{E}_n\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\}$ converges in probability uniformly to $\mathrm{E}\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\}$. In other words, $\exists$ a neighborhood G of $(\mathbf{A}_0, \mathbf{B}_0)$, such that
$$\sup_{(\mathbf{A}_0,\mathbf{B}_0)\in G} \|\mathrm{E}_n\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\} - \mathrm{E}\{\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\}\| \to_P 0;$$

First, check the continuity of the score function. The representation of the function S is given as

$$\mathbf{S}(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y}) = \frac{\partial \log f_{\mathbf{A},\mathbf{B}}(\mathbf{y}, \mathbf{x})}{\partial(\mathbf{A}, \mathbf{B})} = \left(\frac{\partial \log f_{\mathbf{A},\mathbf{B}}(\mathbf{y}, \mathbf{x})}{\partial \mathbf{A}}, \frac{\partial \log f_{\mathbf{A},\mathbf{B}}(\mathbf{y}, \mathbf{x})}{\partial \mathbf{B}} \right).$$

For simplicity, let $\Sigma_{\mathbf{x}}(\mathbf{A}, \mathbf{B}) = \mathbf{A} + \mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}^T(\mathbf{x})\mathbf{B}^T$. It is noted that by the assumption (3.2), $\Sigma_{\mathbf{x}}(\mathbf{A}_0, \mathbf{B}_0)$ is the conditioned covariance matrix of $\mathbf{y}$ given $\mathbf{x}$ at the true $\mathbf{A}_0$ and $\mathbf{B}_0$. This property will be used in the following analysis. For simplicity, denote $\Sigma_{\mathbf{x}}(\mathbf{A}, \mathbf{B}) = \Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}(\mathbf{A}_0, \mathbf{B}_0) = \Sigma_{\mathbf{x}0}$. It follows that

$$\mathbf{S}_1(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y}) = \frac{\partial \log f_{\mathbf{A},\mathbf{B}}(\mathbf{y}, \mathbf{x})}{\partial \mathbf{A}} = \frac{1}{2}\Sigma_{\mathbf{x}}^{-1}(\mathbf{y}\mathbf{y}^T - \Sigma_{\mathbf{x}})\Sigma_{\mathbf{x}}^{-1}, \tag{3.18}$$

$$\mathbf{S}_2(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y}) = \frac{\partial \log f_{\mathbf{A},\mathbf{B}}(\mathbf{y}, \mathbf{x})}{\partial \mathbf{B}} = \Sigma_{\mathbf{x}}^{-1}(\mathbf{y}\mathbf{y}^T - \Sigma_{\mathbf{x}})\Sigma_{\mathbf{x}}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T. \tag{3.19}$$

So $\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})$ is continuous at $(\mathbf{A},\mathbf{B})$. Hence, condition (a) is satisfied.

By the equality $\mathrm{E}\{\mathbf{y}\mathbf{y}^T|\mathbf{x}\} = \Sigma_{\mathbf{x}0}$, (3.18), and (3.19), it is easy to verify that

$$\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})|\mathbf{x}\} = \frac{1}{2}\Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1},$$
$$\mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})|\mathbf{x}\} = \Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T.$$

By the law of total expectation,

$$\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = \frac{1}{2}\mathrm{E}\left\{\Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1}\right\}, \tag{3.20}$$
$$\mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = \mathrm{E}\left\{\Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\right\}. \tag{3.21}$$

Obviously, $\mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} = \mathrm{E}\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} = 0$. Therefore, the condition (b) is also satisfied.

To show that $\mathrm{E}\{S(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ is derivatiable at $(\mathbf{A}_0,\mathbf{B}_0)$, consider that

$$\Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1} = \Sigma_{\mathbf{x}0}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}0}^{-1}+\Delta_x \tag{3.22}$$

with Δ_x a second order term of $\Sigma_{\mathbf{x}}-\Sigma_{\mathbf{x}0}$,

$$\begin{aligned}\Delta_x &= \Sigma_{\mathbf{x}0}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)(\Sigma_{\mathbf{x}}^{-1}-\Sigma_{\mathbf{x}0}^{-1})+(\Sigma_{\mathbf{x}}^{-1}-\Sigma_{\mathbf{x}0}^{-1})\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1}\\ &= \Sigma_{\mathbf{x}0}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}}^{-1}(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}})\Sigma_{\mathbf{x}0}^{-1}+\Sigma_{\mathbf{x}}^{-1}\big(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}}\big)\Sigma_{\mathbf{x}0}^{-1}(\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}})\Sigma_{\mathbf{x}}^{-1}.\end{aligned}$$

Denoting by $\delta_{\mathbf{A}} = \mathbf{A}-\mathbf{A}_0$ and $\delta_{\mathbf{B}} = \mathbf{B}-\mathbf{B}_0$, it follows that

$$\begin{aligned}\Sigma_{\mathbf{x}0}-\Sigma_{\mathbf{x}} &= -\delta_{\mathbf{A}}+\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{B}_0^T-(\mathbf{B}_0+\delta_{\mathbf{B}})\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T(\mathbf{B}_0+\delta_{\mathbf{B}})^T\\ &= -\delta_{\mathbf{A}}-\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\delta_{\mathbf{B}}^T-\delta_{\mathbf{B}}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{B}_0^T-\delta_{\mathbf{B}}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\delta_{\mathbf{B}}^T\\ &= -\mathscr{L}(\delta_{\mathbf{A}},\delta_{\mathbf{B}},\mathbf{x},\mathbf{y})-\delta_{\mathbf{B}}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\delta_{\mathbf{B}}^T,\end{aligned} \tag{3.23}$$

where $\mathscr{L}:(\mathbf{U},\mathbf{V})\to\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})$ is a linear operator defined by

$$\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y}) = \mathbf{U}+\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{V}^T+\mathbf{V}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{B}_0^T. \tag{3.24}$$

Using the above analysis on $\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y}\}$ and $\mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y}\}$ in (3.20-3.21), they can be re-written, respectively, in a sum of two terms with one linear and other second-

order of $\delta_{\mathbf{A}}$ and $\delta_{\mathbf{B}}$,

$$\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = -\frac{1}{2}\mathrm{E}\big\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\delta_{\mathbf{A}},\delta_{\mathbf{B}},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\big\} + \mathbf{D}_1 \tag{3.25}$$

$$\mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = -\mathrm{E}\big\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\delta_{\mathbf{A}},\delta_{\mathbf{B}},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\big\} + \mathbf{D}_2, \tag{3.26}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are two second-order terms of $\delta_{\mathbf{A}}$ and $\delta_{\mathbf{B}}$, which are not of our interest, and hence, omit the details. Therefore, both $\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ and $\mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ are differentiable at $(\mathbf{A}_0,\mathbf{B}_0)$. So the differentiability of $\mathrm{E}\{\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ at $(\mathbf{A}_0,\mathbf{B}_0)$ is shown.

The linear terms in (3.25) and (3.26) define the linear operator in terms of the derivative matrix of $\mathrm{E}\{S(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$. It is easy to verify that the quadratic form of the derivative matrix with respect to the variables $(\mathbf{U},\mathbf{V})$ is given by

$$\begin{aligned} q(\mathbf{U},\mathbf{V}) = & -\frac{1}{2}\mathrm{tr}\Big[\mathbf{U}^T\mathrm{E}\left\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\right\}\Big] \\ & -\mathrm{tr}\Big[\mathbf{V}^T\mathrm{E}\left\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\right\}\Big]. \end{aligned}$$

The negative definiteness of the quadratic form is proven by rewriting

$$\begin{aligned} & -2q(\mathbf{U},\mathbf{V}) \\ &= \mathrm{E}\Big[\mathrm{tr}\big\{\mathbf{U}^T\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\big\} + 2\mathrm{tr}\big\{\mathbf{V}^T\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\big\}\Big] \\ &= \mathrm{E}\Big[\mathrm{tr}\big\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}(\mathbf{U}^T + 2\mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{V}^T)\big\}\Big]. \end{aligned}$$

Using the equality $2\mathrm{tr}(HW) = \mathrm{tr}(H(W+W^T))$ for any matrix W and any symmetric matrix H, it follows that

$$\begin{aligned} & -2q(\mathbf{U},\mathbf{V}) \\ &= \mathrm{E}\Big(\mathrm{tr}\left[\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\left\{\mathbf{U}^T + \mathbf{B}_0\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{V}^T + \mathbf{V}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\mathbf{B}_0^T\right\}\right]\Big) \\ &= \mathrm{E}\Big[\mathrm{tr}\big\{\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\Sigma_{\mathbf{x}0}^{-1}\mathscr{L}(\mathbf{U},\mathbf{V},\mathbf{x},\mathbf{y})\big\}\Big] > 0. \end{aligned}$$

Therefore, the derivative matrix of $\mathrm{E}\{\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ at $(\mathbf{A}_0,\mathbf{B}_0)$ is negative definite, i.e., the condition (c) is true.

Now consider the error $\mathrm{E}_n\{S_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ with

$$\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = \frac{1}{n}\sum_{i=1}^{n} S_1(\mathbf{A},\mathbf{B},\mathbf{x}_i,\mathbf{y}_i).$$

It is represented in a sum of three error terms as follows.

$$\begin{aligned}
&\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} \\
&= \quad \big[\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\big] \\
&\quad + \big[\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\big] \\
&\quad + \big[\mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}\big].
\end{aligned}$$

Obviously, by the strong law of large numbers, the second error term converges in probability to zero,

$$\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} \to_P 0. \tag{3.27}$$

The continuity of $\mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ at $(\mathbf{A}_0,\mathbf{B}_0)$ implies that, for an arbitrary small $\varepsilon > 0$, there exists a neighborhood G_1 of $(\mathbf{A}_0,\mathbf{B}_0)$ in which

$$\|\mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}\| < \varepsilon. \tag{3.28}$$

To estimate the first term, it is represented as

$$\begin{aligned}
&\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} \\
&= \quad \frac{1}{2n}\sum_{i=1}^{n}\Big\{\Sigma_{\mathbf{x}_i}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i})\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i0})\Sigma_{\mathbf{x}_i0}^{-1}\Big\} \\
&= \quad \frac{1}{2n}\sum_{i=1}^{n}\Big[\Sigma_{\mathbf{x}_i}^{-1}(\Sigma_{\mathbf{x}_i0} - \Sigma_{\mathbf{x}_i})\Sigma_{\mathbf{x}_i}^{-1} + \big\{\Sigma_{\mathbf{x}_i}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i0})\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i0})\Sigma_{\mathbf{x}_i0}^{-1}\big\}\Big].
\end{aligned}$$

By (3.23) and (3.24),

$$\|\Sigma_{\mathbf{x}_i0} - \Sigma_{\mathbf{x}_i}\| \le \|\mathbf{A}-\mathbf{A}_0\| + \big(2\|\mathbf{B}_0\| + \|\mathbf{B}-\mathbf{B}_0\|\big)\|\mathbf{B}-\mathbf{B}_0\|\|\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\|.$$

The sequence $\{\|\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\|\}$ is obviously bounded. Hence, for the ε mentioned

above, there exists a neighborhood G_2 of $(\mathbf{A}_0, \mathbf{B}_0)$ such that $\|\Sigma_{\mathbf{x}_i 0} - \Sigma_{\mathbf{x}_i}\| \leq \varepsilon$ for all i. On the other hand, $\|\Sigma_{\mathbf{x}}^{-1}\| \leq \|\mathbf{A}^{-1}\|$ for any $\mathbf{x}$. Let G_3 be a neighborhood of $\mathbf{A}_0$ in which $\|\mathbf{A}^{-1}\| \leq 2\|\mathbf{A}_0^{-1}\|$. It is seen that in that intersected set $G_2 \cap G_3$,

$$\|\Sigma_{\mathbf{x}_i}^{-1}(\Sigma_{\mathbf{x}_i 0} - \Sigma_{\mathbf{x}_i})\Sigma_{\mathbf{x}_i}^{-1}\| \leq \|\Sigma_{\mathbf{x}_i}^{-1}\|\|\Sigma_{\mathbf{x}_i 0} - \Sigma_{\mathbf{x}_i}\|\|\Sigma_{\mathbf{x}_i}^{-1}\| \leq (2\|\mathbf{A}_0^{-1}\|)^2\varepsilon, \quad \forall \mathbf{x}_i. \tag{3.29}$$

It is also observed that

$$\|\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1}\| = \|\Sigma_{\mathbf{x}_i}^{-1}(\Sigma_{\mathbf{x}_i 0} - \Sigma_{\mathbf{x}_i})\Sigma_{\mathbf{x}_i 0}^{-1}\| \leq (2\|\mathbf{A}_0^{-1}\|)^2\varepsilon.$$

Now the error $\Sigma_{\mathbf{x}_i}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})\Sigma_{\mathbf{x}_i 0}^{-1}$ can be estimated as,

$$\begin{aligned}
&\|\Sigma_{\mathbf{x}_i}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})\Sigma_{\mathbf{x}_i 0}^{-1}\| \\
&= \|(\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1})(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})\Sigma_{\mathbf{x}_i}^{-1} + \Sigma_{\mathbf{x}_i 0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0})(\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1})\| \\
&\leq \|\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1}\|\|\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0}\|\|\Sigma_{\mathbf{x}_i}^{-1}\| + \|\Sigma_{\mathbf{x}_i 0}^{-1}\|\|\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0}\|\|\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i 0}^{-1}\| \\
&\leq 2\|\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0}\|(2\|\mathbf{A}_0^{-1}\|)^3\varepsilon \\
&\leq 2C_0(2\|\mathbf{A}_0^{-1}\|)^3\varepsilon,
\end{aligned} \tag{3.30}$$

where C_0 is an upper bound of $\|\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i 0}\|$ for all i. The estimates (3.29) and (3.30) yield

$$\|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0, \mathbf{B}_0, \mathbf{x}, \mathbf{y})\}\| \leq \frac{1}{2}((2\|\mathbf{A}_0^{-1}\|)^2 + C_0(2\|\mathbf{A}_0^{-1}\|)^3)\varepsilon.$$

So it is concluded that there exits a neighborhood G_0 in which

$$\|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0, \mathbf{B}_0, \mathbf{x}, \mathbf{y})\}\| < \varepsilon. \tag{3.31}$$

Combining (3.27), (3.28), and (3.31), the following estimate

$$\begin{aligned}
&\|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A}, \mathbf{B}, \mathbf{x}, \mathbf{y})\| \\
&< 2\varepsilon + \|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0, \mathbf{B}_0, \mathbf{x}, \mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0, \mathbf{B}_0, \mathbf{x}, \mathbf{y})\}\|
\end{aligned}$$

holds in the neighborhood $G = G_0 \cap G_1$ of $(\mathbf{A}_0, \mathbf{B}_0)$. Therefore,

$$\begin{aligned} & \mathrm{P}\Big(\sup_{(\mathbf{A},\mathbf{B})\in G} \|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\| < 3\varepsilon \Big) \\ & \geq \quad \mathrm{P}\Big(\|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\| < \varepsilon \Big) \to 1. \end{aligned}$$

That is,

$$\sup_{(\mathbf{A},\mathbf{B})\in G} \|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\| \to_P 0,$$

Similarly, it can also be shown that there exists a neighborhood G' of $(\mathbf{A}_0, \mathbf{B}_0)$ such that

$$\sup_{(\mathbf{A},\mathbf{B})\in G'} \|\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\| \to_P 0,$$

where $\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = \frac{1}{n}\sum_{i=1}^n \mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x}_i,\mathbf{y}_i)$, because

$$\begin{aligned} & \|\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\| \\ & \leq \quad \big\|\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\big\| \\ & \qquad + \big\|\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\big\| \\ & \qquad + \big\|\mathrm{E}\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}\big\| \end{aligned}$$

and the following results hold for these three terms as (3.27), (3.28), and (3.31) as follows.

(1) $\mathrm{E}_n\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} \to_P 0.$

(2) For an arbitrary small $\varepsilon > 0$, there exists a neighborhood G_1' of $(\mathbf{A}_0, \mathbf{B}_0)$ in which

$$\|\mathrm{E}\{\mathbf{S}_2(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\} - \mathrm{E}\{\mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\| < \varepsilon, \tag{3.32}$$

(3) By (3.30), in the neighborhood $G_2 \cap G_3$,

$$\begin{aligned} & \|\mathrm{E}_n\{\mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}_n\{\mathbf{S}_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\| \\ & \leq \quad \frac{1}{n}\sum_{i=1}^n \big\|\Sigma_{\mathbf{x}_i}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i})\Sigma_{\mathbf{x}_i}^{-1} - \Sigma_{\mathbf{x}_i0}^{-1}(\mathbf{y}_i\mathbf{y}_i^T - \Sigma_{\mathbf{x}_i0})\Sigma_{\mathbf{x}_i0}^{-1}\big\|\|\mathbf{B}\|\big\|\mathbf{S}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\big\| \\ & \leq \quad 2C_0C_1C_2(2\|\mathbf{A}_0^{-1}\|)^3\varepsilon, \end{aligned}$$

where C_1 is a bound of $\|\mathbf{B}\|$ in a neighborhood of $\mathbf{B}_0$ which should be chosen small enough to be contained by $G_2 \cap G_3$, and C_2 is a bound of $\|\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T\|$. Thus, the estimate $\|\mathrm{E}_n\{S_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} - \mathrm{E}_n\{S_1(\mathbf{A}_0,\mathbf{B}_0,\mathbf{x},\mathbf{y})\}\| < \varepsilon$ holds in a neighborhood G_0' of $(\mathbf{A}_0,\mathbf{B}_0)$.

The proof of condition (d) is complete.

In view of the satisfying the condition (a) – (d) in Theorem 3.2.1, the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are shown to be consistent. □

Lemma 3.2.2 (Cramér, Asymptotic Normality)**.** *Let* $\mathbf{z}_1, \mathbf{z}_2, \ldots,$ *be i.i.d. with density* $f_\theta(\mathbf{z})$ *(with respect to* $d\mu$*),* $\theta \in \Theta$*, and let* θ_0 *denote the true value of the parameter, If*

a. Θ *is an open subset of* $\mathbb{R}^k$*;*

b. Second partial derivatives of $\log f_\theta(\mathbf{z})$ *with respect to* θ *exist and are continuous for all* $\mathbf{z}$*, and may be passed under the integral sign in* $\int \log f_\theta(\mathbf{z}) d\mu(\mathbf{z})$*;*

c. There exists a function $M(\mathbf{z})$ *such that* $E_{\theta_0} M(\mathbf{z}) < \infty$ *and each component of the Hessian matrix* $H(\theta,\mathbf{z})$ *of* $\log f_\theta(\mathbf{z})$ *is bounded in absolute value by* $M(\mathbf{z})$ *uniformly in some neighborhood of* θ_0*;*

d. $I(\theta_0) = -E\{H(\theta,\mathbf{z})\}$ *is positive definite;*

e. $f_\theta(\mathbf{z}) = f_{\theta_0}(\mathbf{z})$ *a.e.* $d\mu$ *implies* $\theta = \theta_0$*,*

Then there exists a strongly consistent sequence $\hat{\theta}_n$ *of roots of the likelihood equation such that*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_D N(\mathbf{0}, I^{-1}(\theta_0)).$$

For references of Lemma 3.2.2, see, for example, Cramér (1946) and Lehmann (1998).

Theorem 3.2.2 (Asymptotic Normality)**.** *Under Conditions of Theorem 3.2.1, the maximum likelihood estimators* $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ *satisfy*

$$\sqrt{n}((\hat{\mathbf{A}},\hat{\mathbf{B}})^T - (\mathbf{A}_0,\mathbf{B}_0)^T) \to_D N(\mathbf{0}, I^{-1}\{(\mathbf{A}_0,\mathbf{B}_0)^T\}).$$

Proof. Regarding the asymptotic normality of the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ for the nonparametric regression model (3.5), it remains to check the following condition (a) – (e) for $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ in (3.17). That is, we will check the following conditions one by one.

a. Θ is an open subset of $\mathbb{R}^k$;

b. Second partial derivatives of $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ with respect to $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ exist, are continuous for all $(\mathbf{x},\mathbf{y})$, and may be passed under the integral sign in

$$\int \log f_{\mathbf{A},\mathbf{B}}(\mathbf{z}) d\mu(\mathbf{x},\mathbf{y});$$

c. There exists a function $M(\mathbf{z})$ such that $\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0} M(\mathbf{z}) < \infty$ and each component of the Hessian matrix $H(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})$ of $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ is bounded in absolute value by $M(\mathbf{x},\mathbf{y})$ uniformly in some neighborhood of θ_0;

d. $\mathrm{I}\{(\mathbf{A}_0,\mathbf{B}_0)\} = -\mathrm{E}\{H(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ is positive definite;

e. $f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{A}_0,\mathbf{B}_0}(\mathbf{x},\mathbf{y})$ a.e. $d\mu$ implies $(\mathbf{A},\mathbf{B}) = (\mathbf{A}_0,\mathbf{B}_0)$.

The condition (a) is obviously true. Since $\mathbf{y}|\mathbf{x}$ follows a multivariate normal distribution, with density

$$f_{\mathbf{A},\mathbf{B}}(\mathbf{y},\mathbf{x}) \propto |\Sigma_{\mathbf{x}}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{y}\}$$

and $\Sigma_{\mathbf{x}}$ is a matrix of polynomial functions with respect to the parameters $(\mathbf{A},\mathbf{B})$, the function $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ is a rational function with respect to the parameters $(\mathbf{A},\mathbf{B})$. The existence and continuity for the second partial derivatives of $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ is thus assured. Next, consider the condition (c).

For simplicity. let $\mathbf{F} = \mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T$ and $\mathbf{G} = \mathbf{F}\mathbf{B}^T$. Based on equations (3.18) and (3.19), and using

$$\frac{\partial \Sigma_{\mathbf{x}}}{\partial a_{ij}} = \mathbf{e}_i \mathbf{e}_j^T, \quad \frac{\partial \Sigma_{\mathbf{x}}}{\partial b_{st}} = \frac{\partial \mathbf{G}}{\partial b_{st}} = \mathbf{e}_s \mathbf{e}_t^T \mathbf{F}^T + \mathbf{F}\mathbf{e}_t \mathbf{e}_s^T, \quad \frac{\partial \mathbf{F}}{\partial b_{st}} = \mathbf{e}_s \mathbf{e}_t^T \mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T,$$

it follows that

$$\frac{\partial \mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial a_{ij}} = -\frac{1}{2}\Sigma_{\mathbf{x}}^{-1}(\mathbf{e}_i \mathbf{e}_j^T \Sigma_{\mathbf{x}}^{-1} \mathbf{y}\mathbf{y}^T + \mathbf{y}\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{e}_i \mathbf{e}_j + \mathbf{e}_i \mathbf{e}_j^T)\Sigma_{\mathbf{x}}^{-1}, \tag{3.33}$$

$$\frac{\partial \mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial b_{st}} = -\frac{1}{2}\Sigma_{\mathbf{x}}^{-1}\left(\frac{\partial \mathbf{G}}{\partial b_{st}}\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\mathbf{y}^T + \mathbf{y}\mathbf{y}^T\Sigma_x^{-1}\frac{\partial \mathbf{G}}{\partial b_{st}} + \frac{\partial \mathbf{G}}{\partial b_{st}}\right)\Sigma_{\mathbf{x}}^{-1}, \tag{3.34}$$

$$\frac{\partial \mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial a_{ij}} = -\Sigma_{\mathbf{x}}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\mathbf{y}^T + \mathbf{y}\mathbf{y}^T\Sigma_{\mathbf{x}}^{-1}\mathbf{e}_i\mathbf{e}_j + \mathbf{e}_i\mathbf{e}_j^T)\Sigma_{\mathbf{x}}^{-1}\mathbf{F}, \tag{3.35}$$

$$\begin{aligned}\frac{\partial \mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial b_{st}} = &-\Sigma_{\mathbf{x}}^{-1}\left(\frac{\partial \mathbf{C}}{\partial b_{st}}\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\mathbf{y}^T + \mathbf{y}\mathbf{y}^T\Sigma_x^{-1}\frac{\partial \mathbf{C}}{\partial b_{st}} + \frac{\partial \mathbf{C}}{\partial b_{st}}\right)\Sigma_{\mathbf{x}}^{-1}\mathbf{F} \\ &+ (\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\mathbf{y}^T\Sigma_{\mathbf{x}}^{-1} - \Sigma_{\mathbf{x}}^{-1})\frac{\partial \mathbf{F}}{\partial b_{st}}.\end{aligned} \tag{3.36}$$

Let M_1 be the neighborhood of $\mathbf{A}_0$, such that $\|\mathbf{A}^{-1}\| \le 2\|\mathbf{A}_0^{-1}\|$. We have $\|\Sigma_{\mathbf{x}}^{-1}\| \le \|\mathbf{A}^{-1}\| \le 2\|\mathbf{A}_0^{-1}\|$ for all x. Let M_2 be the neighborhood of $\mathbf{B}_0$, such that $\|\mathbf{B}\| \le 2\|\mathbf{B}_0\|$. Then, in the neighborhood $M = M_1 \cap M_2$ of θ_0,

$$\begin{aligned}\|\frac{\partial \mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial a_{ij}}\| &\le \frac{1}{2}\|\Sigma_{\mathbf{x}}^{-1}\|^2\big(\|\mathbf{e}_i\mathbf{e}_j^T\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\mathbf{y}^T\| + \|\mathbf{y}\mathbf{y}^T\Sigma_{\mathbf{x}}^{-1}\mathbf{e}_i\mathbf{e}_j^T\| + \|\mathbf{e}_i\mathbf{e}_j^T\|\big) \\ &\le \frac{1}{2}\|\Sigma_{\mathbf{x}}^{-1}\|^2\big(2\|\Sigma_{\mathbf{x}}^{-1}\|\|\mathbf{y}\mathbf{y}^T\| + 1\big) \\ &\le 2\|\mathbf{A}_0^{-1}\|^2\big(4\|\mathbf{A}_0^{-1}\|\|\mathbf{y}\mathbf{y}^T\| + 1\big) =: M_1.\end{aligned}$$

Similarly,

$$\begin{aligned}\|\frac{\partial \mathbf{S}_1(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial b_{st}}\| &\le 2M_1\|\mathbf{F}\| \le 4M_1\|\mathbf{B}_0\|\|\mathbf{S}(\mathbf{x})\|^2 =: M_2, \\ \|\frac{\partial \mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial a_{ij}}\| &\le M_2, \\ \|\frac{\partial \mathbf{S}_2(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial b_{st}}\| &\le 4M_2\|\mathbf{B}_0\|\|\mathbf{S}(\mathbf{x})\|^2 + 2\|\mathbf{A}_0^{-1}\|\big(2\|\mathbf{A}_0^{-1}\|\|\mathbf{y}\mathbf{y}^T\| + 1\big)\|\mathbf{S}(\mathbf{x})\|^2 \\ &=: M_3.\end{aligned}$$

Thus, each second derivative of $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ is bounded by

$$M(\mathbf{x},\mathbf{y}) = \max(M_1, M_2, M_3).$$

It is noticed that

$$\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(\|\mathbf{y}\mathbf{y}^T\|) = \mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}\{\mathrm{tr}(\mathbf{y}\mathbf{y}^T)\} = \mathrm{tr}\left\{\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(\mathbf{y}\mathbf{y}^T)\right\} = \mathrm{tr}\left\{\mathrm{E}(\Sigma_{\mathbf{x}0})\right\}.$$

Hence,

$$\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(M_1(\mathbf{x},\mathbf{y})) = 2\|\mathbf{A}_0^{-1}\|^2\left[4\|\mathbf{A}_0^{-1}\mathrm{tr}\left\{\mathrm{E}(\Sigma_{\mathbf{x}0})\right\}+1\right] < \infty,$$

In the same fashion, it follows that in the neighborhood of $(\mathbf{A}_0,\mathbf{B}_0)$, $\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(M_2)$ and $\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(M_3)$ are finite, too, concluding that $\mathrm{E}_{\mathbf{A}_0,\mathbf{B}_0}(M(\mathbf{x},\mathbf{y})) < \infty$. Hence, condition (c) is true.

The condition (c) implies that $\mathrm{E}\left\{\frac{\partial \mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial(\mathbf{A},\mathbf{B})}\right\}$ converges uniformly in the neighborhood of $(\mathbf{A}_0,\mathbf{B}_0)$. Therefore, $\frac{\partial}{\partial(\mathbf{A},\mathbf{B})}\mathrm{E}\{\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\} = \mathrm{E}\left\{\frac{\partial \mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial(\mathbf{A},\mathbf{B})}\right\}$. Accordingly, it has been shown that second partial derivatives of $\log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})$ with respect to $(\mathbf{A},\mathbf{B})$ exist and are continuous for all $(\mathbf{x},\mathbf{y})$, and may be passed under the integral sign in $\int \log f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y})d\mu(\mathbf{x},\mathbf{y})$, which completes the prove of condition (b).

It has been shown that $\frac{\partial}{\partial(\mathbf{A},\mathbf{B})}\mathrm{E}\{\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ is negative definite. By condition (b), $-\mathrm{E}\{\frac{\partial \mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})}{\partial(\mathbf{A},\mathbf{B})}\} = -\frac{\partial}{\partial(\mathbf{A},\mathbf{B})}\mathrm{E}\{\mathbf{S}(\mathbf{A},\mathbf{B},\mathbf{x},\mathbf{y})\}$ is positive definite. Condition (d) is satisfied.

Finally, the condition $f_{\mathbf{A},\mathbf{B}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{A}_0,\mathbf{B}_0}(\mathbf{x},\mathbf{y})$ a.e. $d\mu$ for all $\mathbf{z} = (\mathbf{x},\mathbf{y}) \in \mathbb{R}^p$ is equivalent to

$$|\Sigma_{\mathbf{x}}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma_{\mathbf{x}}^{-1}\mathbf{y}\right) = |\Sigma_{\mathbf{x}0}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma_{\mathbf{x}0}^{-1}\mathbf{y}\right) \text{ a.e. } d\mu, \ \forall \mathbf{x},\mathbf{y},$$

which is equivalent to

$$|\Sigma_{\mathbf{x}0}|^{\frac{1}{2}}|\Sigma_{\mathbf{x}}|^{-\frac{1}{2}} = \exp\left\{-\frac{1}{2}\mathbf{y}^T(\Sigma_{\mathbf{x}0}^{-1}-\Sigma_{\mathbf{x}}^{-1})\mathbf{y}\right\} \text{ a.e. } d\mu, \ \forall \mathbf{x},\mathbf{y}.$$

The left side of the above equation is a constant for fixed $\mathbf{x}$, thus

$$\mathbf{y}^T(\Sigma_{\mathbf{x}0}^{-1}-\Sigma_{\mathbf{x}}^{-1})\mathbf{y} = c(\mathbf{x}) \text{ a.e. } d\mu, \ \forall \mathbf{x},\mathbf{y}.$$

where $c(\mathbf{x}) = \log|\Sigma_{\mathbf{x}}| - \log|\Sigma_{\mathbf{x}0}|$ depends on $\mathbf{x}$ only. Next, it is shown that $c(\mathbf{x}) = 0$ a.e. $d\mu_{\mathbf{x}}$ with the marginal density $\mu_{\mathbf{x}}$. It is proceeded by contradiction. Let $D = \{\mathbf{x} \mid c(\mathbf{x}) \neq 0\}$. If $\mu_{\mathbf{x}}(D) \neq 0$, then D has an open subset D_0. Let

$$\Omega = \left\{(\mathbf{x},\mathbf{y}) \mid \mathbf{y}^T(\Sigma_{\mathbf{x}0}^{-1}-\Sigma_{\mathbf{x}}^{-1})\mathbf{y} = c(\mathbf{x}), \ \mathbf{x} \in D_0\right\}.$$

Obviously, $\mu(\Omega) > 0$ since $\mu_{\mathbf{x}}(D_0) > 0$, which implies that Ω has an open set Ω_0. Hence, one can find two points $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}, a\mathbf{y})$ in $\Omega_0 \subset \Omega$ with $|a| \neq 1$. It follows that

$$a^2 c(\mathbf{x}) = a^2 \mathbf{y}^T (\Sigma_{\mathbf{x}0}^{-1} - \Sigma_{\mathbf{x}}^{-1})\mathbf{y} = (a\mathbf{y})^T (\Sigma_{\mathbf{x}0}^{-1} - \Sigma_{\mathbf{x}}^{-1})(a\mathbf{y}) = c(\mathbf{x}),$$

which yields $c(\mathbf{x}) = 0$ in contradiction to that $\mathbf{x} \in D_0 \subset D$. Therefore, it is proved that $c(\mathbf{x}) = 0$ a.e. $d\mu_{\mathbf{x}}$, or equivalently, $\Sigma_{\mathbf{x}0} = \Sigma_{\mathbf{x}}$ a.e. $d\mu_{\mathbf{x}}$. Since $\Sigma_{\mathbf{x}}$ is a one-to-one function from $(\mathbf{A}, \mathbf{B})$ to $\Sigma_{\mathbf{x}}$, it is concluded that $(\mathbf{A}, \mathbf{B}) = (\mathbf{A}_0, \mathbf{B}_0)$ and complete the proof of condition (e).

It is known that there is a unique root of the likelihood equation for every n, which is the maximum likelihood estimator $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. Moreover, it is shown the consistency of the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. By Theorem 3.2.2 accordingly, the maximum likelihood estimators $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ have asymptotic normality. □

3.3 Covariance Regression Model II

Model (3.1) assumes that all the variables $\mathbf{x}_i$ are continuous. In many real applications, categorical variables such as gender may also affect the observations. In this section, an extension of the nonparametric covariance regression model is considered by taking the potential categorical variables into account.

Let $\mathbf{z} \in \mathbb{R}^r$ be a dummy variable that represent categorical predicting variables that have $0/1$ elements. The extended covariance regression model of (3.1) has the following form

$$\Sigma_{\mathbf{x}} = \mathbf{A} + \mathbf{B}\mathbf{S}(\mathbf{x})\mathbf{S}(\mathbf{x})^T \mathbf{B}^T + \mathbf{C}\mathbf{z}\mathbf{z}^T \mathbf{C}^T, \tag{3.37}$$

where $\mathbf{S}(\mathbf{x})$, $\mathbf{A}$, and $\mathbf{B}$ are defined as before, and $\mathbf{C}$ is a $p \times r$ matrix that needs to be estimated as the unknown matrices $\mathbf{A}$ and $\mathbf{B}$.

To estimate the parameter matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, consider the nonparametric regression model:

$$\mathbf{y}_i \;=\; \boldsymbol{\mu}_{\mathbf{x}_i, \mathbf{z}_i} + \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i + \tau_i \mathbf{C}\mathbf{z}_i + \boldsymbol{\varepsilon}_i, \tag{3.38}$$

based on the set of random samples $\{\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, i = 1, 2, \ldots, n\}$ from population $\{\mathbf{y}, \mathbf{x}, \mathbf{z}\}$,

where γ_i and τ_i are random variables with the normalization assumptions:

$$\begin{aligned}
\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n &\sim N_d(\mathbf{0},\mathbf{I}_{d\times d}),\\
\tau_1,\ldots,\tau_n &\sim N(0,1),\ \ \tau_i \text{ and } \boldsymbol{\gamma}_i \text{ independent, } \sqrt[n]{}\\
\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n &\sim N_p(\mathbf{0},\mathbf{A}_{p\times p}).
\end{aligned} \tag{3.39}$$

It is assumed that the conditioned mean $\boldsymbol{\mu}_{\mathbf{x}_i,\mathbf{z}_i}$ given $\mathbf{x}_i$ and $\mathbf{z}_i$, is known for $i = 1,\ldots,n$.

The above assumptions result in the conditional normal distribution of $\{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n\}$ and $\{\tau_1,\ldots,\tau_n\}$ given $\{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{Y},\mathbf{X},\mathbf{Z}\}$, which the theoretical derivation can be described as follows.

First, it follow that

$$\begin{aligned}
(\mathbf{y}_i \mid \boldsymbol{\gamma}_i,\tau_i,\mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C}) &\sim N_p(\boldsymbol{\mu}_{\mathbf{x}_i,\mathbf{z}_i}+\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i+\tau_i\mathbf{C}\mathbf{z}_i,\mathbf{A})\\
\boldsymbol{\gamma}_i &\sim N_d(\mathbf{0},\mathbf{I}_{d\times d})\\
\tau_i &\sim N(0,1),\ \boldsymbol{\gamma}_i \text{ and } \tau_i \text{ are independent,}
\end{aligned}$$

then the conditional distribution of $\{\mathbf{y}_i,\boldsymbol{\gamma}_i,\tau_i\}$ given $\{\mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C}\}$ can be expressed as

$$\begin{aligned}
&f(\mathbf{y}_i,\boldsymbol{\gamma}_i,\tau_i \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})\\
&= (2\pi)^{-\frac{p+d+1}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\big[-\frac{1}{2}\boldsymbol{\gamma}_i^T\boldsymbol{\gamma}_i-\frac{1}{2}\tau_i^2\\
&\quad -\frac{1}{2}\left\{\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i,\mathbf{z}_i}-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i,\mathbf{z}_i}-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}\big]
\end{aligned}$$

so the conditional distribution of $(\mathbf{y}_i,\boldsymbol{\gamma}_i)$ given $(\mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})$ is known as

$$\begin{aligned}
&f(\mathbf{y}_i,\boldsymbol{\gamma}_i \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})\\
&= \int f(\mathbf{y}_i,\boldsymbol{\gamma}_i,\tau_i \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})d\tau_i\\
&= \int(2\pi)^{-\frac{p+d+1}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\big[\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{e}_i-2\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i+\\
&\quad \boldsymbol{\gamma}_i^T\left\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\right\}\boldsymbol{\gamma}_i-\sigma_1^2(\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{e}_i-\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i)^2\big]\Big)\\
&\quad \exp\big\{-\frac{1}{2\sigma_1^2}(\tau_i-\mu_1)^2\big\}d\tau_i
\end{aligned}$$

$$
\begin{aligned}
= \quad & (2\pi)^{-\frac{p+d}{2}}|\mathbf{A}|^{-\frac{1}{2}}(\sigma_1^2)^{\frac{1}{2}}\exp\Big(-\frac{1}{2}\big[\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{e}_i - 2\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{e}_i + \\
& \boldsymbol{\gamma}_i^T\left\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\right\}\boldsymbol{\gamma}_i - \sigma_1^2(\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{e}_i - \mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i)^2\big]\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_1^2 &= (1+\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\mathbf{z}_i)^{-1}, \\
\mu_1 &= \sigma_1^2\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i - \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i\right\}.
\end{aligned}
$$

So the conditional distribution of $\boldsymbol{\gamma}_i$ given $(\mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{B},\mathbf{C})$ can be derived as follows,

$$
\begin{aligned}
& f(\boldsymbol{\gamma}_i \mid \mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{B},\mathbf{C}) \\
& \quad \propto \quad f(\mathbf{y}_i,\boldsymbol{\gamma}_i \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C}) \\
& \quad \propto \quad \exp\Big(-\frac{1}{2}\big[\boldsymbol{\gamma}_i^T\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{B}\mathbf{S}(\mathbf{x}_i)\}\boldsymbol{\gamma}_i - \\
& \qquad 2\boldsymbol{\gamma}_i^T\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{e}_i\big]\Big) \\
& \quad \propto \quad \exp\big\{-\frac{1}{2}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_{\boldsymbol{\gamma}_i})^T\Sigma_{\boldsymbol{\gamma}_i}^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_{\boldsymbol{\gamma}_i})\big\},
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma_{\boldsymbol{\gamma}_i} &= \{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{B}\mathbf{S}(\mathbf{x}_i)\}^{-1}, \\
\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} &= \Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{e}_i.
\end{aligned}
$$

Therefore, the conditional distribution of $\{\gamma_1,\ldots,\gamma_n\}$ given $\{\mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{B},\mathbf{C}\}$ is as follows:

$$
\begin{aligned}
(\boldsymbol{\gamma}_i|\mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{B},\mathbf{C}) &\sim N_d(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\Sigma_{\boldsymbol{\gamma}_i}), \text{ where} \\
\Sigma_{\boldsymbol{\gamma}_i} &= \{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{B}\mathbf{S}(\mathbf{x}_i)\}^{-1}, \\
\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} &= \Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{e}_i.
\end{aligned}
$$

Under the similar fashion, the conditional distribution of $(\mathbf{y}_i,\tau_i)$ given $(\mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})$ is derived as

$f(\mathbf{y}_i,\tau_i \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{A},\mathbf{B},\mathbf{C})$

$$
\begin{aligned}
&= \int f(\mathbf{y}_i, \boldsymbol{\gamma}_i, \tau_i \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{A}, \mathbf{B}, \mathbf{C}) d\boldsymbol{\gamma}_i \\
&= \int (2\pi)^{-\frac{p+d+1}{2}} |\mathbf{A}|^{-\frac{1}{2}} \exp\Big[-\frac{1}{2}\big\{ \mathbf{e}_i^T \mathbf{A}^{-1} \mathbf{e}_i + \tau_i^2 (1 + \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C} \mathbf{z}_i) - 2\tau_i \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \mathbf{e}_i \\
&\quad -(\mathbf{e}_i - \tau_i \mathbf{C}\mathbf{z}_i)^T \mathbf{A}^{-1} \mathbf{B} \mathbf{S}(\mathbf{x}_i) \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} (\mathbf{e}_i - \tau_i \mathbf{C}\mathbf{z}_i) \big\} \Big] \\
&\quad \exp\big\{ -\frac{1}{2\sigma_1^2} (\boldsymbol{\gamma}_i - \boldsymbol{\mu}_2)^T \Sigma_2^{-1} (\boldsymbol{\gamma}_i - \boldsymbol{\mu}_2) \big\} d\boldsymbol{\gamma}_i \\
&= (2\pi)^{-\frac{p+1}{2}} |\mathbf{A}|^{-\frac{1}{2}} |\Sigma_2|^{\frac{1}{2}} \exp\Big[-\frac{1}{2}\big\{ \mathbf{e}_i^T \mathbf{A}^{-1} \mathbf{e}_i + \tau_i^2 (1 + \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C} \mathbf{z}_i) - \\
&\quad 2\tau_i \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \mathbf{e}_i - \boldsymbol{\mu}_2^T \Sigma_2^{-1} \boldsymbol{\mu}_2 \big\} \Big],
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma_2 &= \{\mathbf{I} + \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B} \mathbf{S}(\mathbf{x}_i)\}^{-1}, \\
\boldsymbol{\mu}_2 &= \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} (\mathbf{e}_i - \tau_i \mathbf{C}\mathbf{z}_i).
\end{aligned}
$$

So the conditional distribution of τ_i given $\{\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}$ can be derived as follows,

$$
\begin{aligned}
&f(\tau_i \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}) \\
&\propto f(\mathbf{y}_i, \tau_i \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{A}, \mathbf{B}, \mathbf{C}) \\
&\propto \exp\Big\{ -\frac{1}{2}\Big(\tau_i^2 \left[1 + \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \left\{ \mathbf{I} - \mathbf{B}\mathbf{S}(\mathbf{x}_i) \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} \right\} \mathbf{C}\mathbf{z}_i \right] - \\
&\quad 2\tau_i \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \left\{ \mathbf{I} - \mathbf{B}\mathbf{S}(\mathbf{x}_i) \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} \right\} \mathbf{e}_i \Big) \Big\} \\
&\propto \exp\big\{ -\frac{1}{2\sigma_{\tau_i}^2} (\tau_i - \mu_{\tau_i})^2 \big\},
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_{\tau_i}^2 &= \{1 + \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} (\mathbf{I} - \mathbf{B}\mathbf{S}(\mathbf{x}_i) \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1}) \mathbf{C}\mathbf{z}_i \}^{-1}, \\
\mu_{\tau_i} &= \Sigma_{\tau_i} \mathbf{z}_i^T \mathbf{C}^T \mathbf{A}^{-1} \left\{ \mathbf{I} - \mathbf{B}\mathbf{S}(\mathbf{x}_i) \Sigma_2 \mathbf{S}(\mathbf{x}_i)^T \mathbf{B}^T \mathbf{A}^{-1} \right\} \mathbf{e}_i.
\end{aligned}
$$

Therefore, the conditional distribution of $\{\tau_1, \ldots, \tau_n\}$ given $\{\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is as follows:

$$
(\tau_i | \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}) \sim N_d(\mu_{\tau_i}, \sigma_{\tau_i}^2), \text{ where}
$$

$$
\begin{aligned}
\sigma^2_{\tau_i} &= \left[1+\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\left\{\mathbf{I}-\mathbf{BS}(\mathbf{x}_i)\Sigma_2\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\right\}\mathbf{C}\mathbf{z}_i\right]^{-1},\\
\mu_{\tau_i} &= \sigma^2_{\tau_i}\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\left\{\mathbf{I}-\mathbf{BS}(\mathbf{x}_i)\Sigma_2\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\right\}\mathbf{e}_i, \qquad (3.40)\\
(\boldsymbol{\gamma}_i|\mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{A},\mathbf{B},\mathbf{C}) &\sim N_d(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i},\Sigma_{\boldsymbol{\gamma}_i}),\ \text{where}\\
\Sigma_{\boldsymbol{\gamma}_i} &= \{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{BS}(\mathbf{x}_i)\}^{-1},\\
\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} &= \Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}(\mathbf{I}-\sigma_1^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1})\mathbf{e}_i, \qquad (3.41)
\end{aligned}
$$

with $\Sigma_2=\{\mathbf{I}+\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{BS}(\mathbf{x}_i)\}^{-1}$ and $\sigma_1^2=(1+\mathbf{z}_i^T\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\mathbf{z}_i)^{-1}$.

Similar as in the previous model, the log-likelihood of the parameters based on $\mathbf{X}$, $\mathbf{Z}$, and the matrix $\mathbf{E}=(\mathbf{e}_1^T,\cdots,\mathbf{e}_n^T)^T$ of residuals $\mathbf{e_i}=\mathbf{y}_i-\boldsymbol{\mu}_{\mathbf{x}_i,\mathbf{z}_i}$, $i=1,2,\ldots,n$ is

$$
\begin{aligned}
\ell(\mathbf{A},\mathbf{B},\mathbf{C}:\mathbf{E},\mathbf{X},\mathbf{Z}) &= c-\frac{1}{2}\sum_{i=1}^n\log|\mathbf{A}+\mathbf{BS}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T+\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T|\\
&\quad -\frac{1}{2}\sum_{i=1}^n\mathrm{tr}\left[\left\{\mathbf{A}+\mathbf{BS}(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T+\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\right\}^{-1}\mathbf{e}_i\mathbf{e}_i^T\right]. \qquad (3.42)
\end{aligned}
$$

The data log-likelihood $\ell(\mathbf{A},\mathbf{B},\mathbf{C})$ given $\{\mathbf{Y},\mathbf{X},\mathbf{Z},\mu,\gamma\}$ is

$$
\begin{aligned}
&\ell(\mathbf{A},\mathbf{B},\mathbf{C})\\
&= -\frac{1}{2}\Big[np\log(2\pi)+n\log|\mathbf{A}|+\\
&\qquad +\sum_{i=1}^n\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}\Big]. \qquad (3.43)
\end{aligned}
$$

Therefor, given current estimates $(\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}})$ of $(\mathbf{A},\mathbf{B},\mathbf{C})$, the estimation step of the EM-algorithm is to compute $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$, $\Sigma_{\boldsymbol{\gamma}_i}$, μ_{τ_i}, and $\sigma^2_{\tau_i}$ as follows:

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}&=\mathrm{E}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}},\mathbf{e}_i,\mathbf{x}_i,\mathbf{z}_i), \quad \Sigma_{\boldsymbol{\gamma}_i}=\mathrm{Var}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i,\mathbf{x}_i,\mathbf{z}_i),\\
\mu_{\tau_i}&=\mathrm{E}(\tau_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}},\mathbf{e}_i,\mathbf{x}_i,\mathbf{z}_i), \quad \sigma^2_{\tau_i}=\mathrm{Var}(\tau_i|\hat{\mathbf{A}},\hat{\mathbf{B}},\mathbf{e}_i,\mathbf{x}_i,\mathbf{z}_i).
\end{aligned}
$$

Substitude $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$, $\Sigma_{\boldsymbol{\gamma}_i}$, μ_{τ_i} and $\sigma^2_{\tau_i}$ into the complete data log-likelihood (3.43), it follows that

$$
\mathrm{E}\{\ell(\mathbf{A},\mathbf{B},\mathbf{C})|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}\}
$$

$$
\begin{aligned}
= \quad & -\frac{1}{2}\Big\{np\log(2\pi)+n\log|\mathbf{A}|+ \\
& +\sum_{i=1}^{n}\mathrm{E}\left(\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}\right)\Big\}.
\end{aligned}
$$

By the factorization $\Sigma_{r_i}=\mathbf{K}_i\mathbf{K}_i^T$, the last term can be simplified as shown below.

$$
\begin{aligned}
& \mathrm{E}\left[\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\gamma}_i-\tau_i\mathbf{C}\mathbf{z}_i\right\}|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}\right] \\
= \quad & \left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i\right\} \\
& +\mathrm{tr}\left[\mathbf{A}^{-1}\left\{\mathbf{BS}(\mathbf{x}_i)\Sigma_{r_i}\mathbf{S}(\mathbf{x}_i)^T\mathbf{B}^T+\sigma_{\tau_i}^2\mathbf{C}\mathbf{z}_i\mathbf{z}_i^T\mathbf{C}^T\right\}\right] \\
= \quad & \left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i\right\} \\
& +\mathrm{tr}\left[\left\{\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right\}^T\mathbf{A}^{-1}\left\{\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i\right\}+(\sigma_{\tau_i}\mathbf{C}\mathbf{z}_i)^T\mathbf{A}^{-1}(\sigma_{\tau_i}\mathbf{C}\mathbf{z}_i)\right] \\
= \quad & \mathrm{tr}\Big[\left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i,\ -\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i,\ -\sigma_{\tau_i}\mathbf{C}\mathbf{z}_i\right\}^T\mathbf{A}^{-1} \\
& \qquad \left\{\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i,\ -\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i,\ -\sigma_{\tau_i}\mathbf{C}\mathbf{z}_i\right\}\Big] \\
= \quad & \mathrm{tr}\Big\{\left(\mathbf{E}_i^*-\mathbf{D}\mathbf{X}_i^*\right)^T\mathbf{A}^{-1}\left(\mathbf{E}_i^*-\mathbf{D}\mathbf{X}_i^*\right)\Big\}.
\end{aligned}
$$

Here the following representation was used.

$$
\begin{aligned}
& \left(\mathbf{e}_i-\mathbf{BS}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}-\mu_{\tau_i}\mathbf{C}\mathbf{z}_i,\ -\mathbf{BS}(\mathbf{x}_i)\mathbf{K}_i,\ -\sigma_{\tau_i}\mathbf{C}\mathbf{z}_i\right) \\
= \quad & (\mathbf{e}_i,\mathbf{0},0)-(\mathbf{B},\mathbf{C})\begin{pmatrix}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} & \mathbf{S}(\mathbf{x}_i)\mathbf{K}_i & 0\\ \mu_{\tau_i}\mathbf{z}_i & 0 & \sigma_{\tau_i}\mathbf{z}_i\end{pmatrix}=\mathbf{E}_i^*-\mathbf{D}\mathbf{X}_i^*,
\end{aligned}
$$

with $\mathbf{E}_i^*=(\mathbf{e}_i,\mathbf{0},0)$, $\mathbf{D}=(\mathbf{B},\mathbf{C})$, and $\mathbf{X}_i^*=\begin{pmatrix}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} & \mathbf{S}(\mathbf{x}_i)\mathbf{K}_i & 0\\ \mu_{\tau_i}\mathbf{z}_i & 0 & \sigma_{\tau_i}\mathbf{z}_i\end{pmatrix}$. Therefore, denoting $\mathbf{X}^*=(\mathbf{X}_1^*,\ldots,\mathbf{X}_n^*)$ and $\mathbf{E}^*=(\mathbf{E}_1^*,\ldots\mathbf{E}_n^*)$, the expected value of the complete data log-likelihood can be expressed as

$$
\begin{aligned}
& \mathrm{E}\{\ell(\mathbf{A},\mathbf{B},\mathbf{C})|\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}\} \\
= \quad & -\frac{1}{2}\left[np\log(2\pi)+n\log|\mathbf{A}|+\mathrm{tr}\Big\{\left(\mathbf{E}^*-\mathbf{D}\mathbf{X}^*\right)^T\mathbf{A}^{-1}\left(\mathbf{E}^*-\mathbf{D}\mathbf{X}^*\right)\Big\}\right],
\end{aligned}
$$

yielding the expected likelihood for multivariate normal regression whose maximizers

are given by

$$\tilde{\mathbf{D}} = \mathbf{E}^*\mathbf{X}^{*T}(\mathbf{X}^*\mathbf{X}^{*T})^{-1}, \tag{3.44}$$

$$\tilde{\mathbf{A}} = (\mathbf{E}^* - \hat{\mathbf{D}}\mathbf{X}^*)(\mathbf{E}^* - \hat{\mathbf{D}}\mathbf{X}^*)^T/n. \tag{3.45}$$

The above EM formula can also be simplified by using

$$\mathbf{w}_i = \begin{pmatrix} \mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} \\ \mathbf{z}_i\mu_{\tau_i} \end{pmatrix}, \ \mathbf{Q}_i = \begin{pmatrix} \mathbf{S}(\mathbf{x}_i)\Sigma_{\boldsymbol{\gamma}_i}\mathbf{S}(\mathbf{x}_i)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_i\mathbf{z}_i^T\sigma^2_{\boldsymbol{\gamma}_i} \end{pmatrix}, \ \mathbf{W} = (\mathbf{w}_i, \ldots, \mathbf{w}_n).$$

It produces the equalities

$$\begin{aligned} \mathbf{E}^*\mathbf{X}^{*T} &= \mathbf{E}\mathbf{W}, \\ \mathbf{X}^*\mathbf{X}^{*T} &= \mathbf{W}\mathbf{W}^T + \sum_{i=1}^{n}\mathbf{Q}_i, \\ (\mathbf{E}^* - \hat{\mathbf{D}}\mathbf{X}^*)(\mathbf{E}^* - \hat{\mathbf{D}}\mathbf{X}^*)^T &= (\mathbf{E} - \hat{\mathbf{D}}\mathbf{W})(\mathbf{E} - \hat{\mathbf{D}}\mathbf{W})^T + \hat{\mathbf{D}}\big(\sum_{i=1}^{n}\mathbf{Q}_i\big)\hat{\mathbf{D}}^T. \end{aligned}$$

Let $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ be the current estimation, the EM algorithm for estimating the parameters in model (3.38) can be explained as follows.

EM algorithm for the regression model (3.38).

1. Estimate $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} = \mathrm{E}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \mathbf{e}_i, \mathbf{x}_i, \mathbf{z}_i)$, $\Sigma_{\boldsymbol{\gamma}_i} = \mathrm{Var}(\boldsymbol{\gamma}_i|\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \mathbf{e}_i, \mathbf{x}_i, \mathbf{z}_i)$, $\mu_{\tau_i} = \mathrm{E}(\tau_i|\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \mathbf{e}_i, \mathbf{x}_i, \mathbf{z}_i)$, and $\sigma^2_{\tau_i} = \mathrm{Var}(\tau_i|\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \mathbf{e}_i, \mathbf{x}_i, \mathbf{z}_i)$.

2. Compute $\mathbf{w}_i = \begin{pmatrix} \mathbf{S}(\mathbf{x}_i)\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} \\ \mathbf{z}_i\mu_{\tau_i} \end{pmatrix}$, $\mathbf{T} = \mathbf{E} - \hat{\mathbf{D}}\mathbf{W}$, and $\mathbf{Q} = \sum_{i=1}^{n}\mathbf{Q}_i$.

3. Update $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$ by

$$\hat{\mathbf{A}} = \big(\mathbf{T}\mathbf{T}^T + \hat{\mathbf{D}}\mathbf{Q}\hat{\mathbf{D}}^T\big)/n, \quad \hat{\mathbf{D}} = \mathbf{E}\mathbf{W}^T(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \mathbf{Q})^{-\mathbf{1}}.$$

This procedure is repeated until a desired convergence criterion has been satisfied. Since $\hat{\mathbf{D}} = (\hat{\mathbf{B}}, \ \hat{\mathbf{C}})$, namely, $\hat{\mathbf{B}}$ is the first q columns of $\hat{\mathbf{D}}$ and $\hat{\mathbf{C}}$ is the last r columns of $\hat{\mathbf{D}}$, the estimates of $\mathbf{B}$ and $\mathbf{C}$ can be obtained.

3.4 Numerical Studies

3.4.1 Simulation Studies

Example 1

First, consider a simulation study for covariance regression model (3.1). The simulated data sets are generated as follows. Let $\mathbf{A}$ and $\mathbf{B}$ be

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0.8 & -0.2 & -0.2 \\ -0.2 & 0.8 & -0.2 \\ -0.2 & -0.2 & 0.8 \\ -0.2 & -0.2 & -0.2 \end{bmatrix},$$

where $\mathbf{A}$ is a 4×4 matrix and $\mathbf{B}$ is a 4×3 matrix (i.e., $p = 4$ and $q = 3$).

First, independently generate each element of $\mathbf{X}$, x_{ij} from Uniform distribution $U(0,1)$ for $i = 1, \ldots, n,\ j = 1, \ldots, 3$. The random error vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are taken from a multivariate normal population with zero mean and covariance matrix $\mathbf{A}$. Five knots (i.e., $J = 5$), which denoted as $\kappa_1, \ldots, \kappa_5$, are selected at equal intervals over the range of $\mathbf{X}$ (i.e., [0,1]). Hence, the smoothing spline $\mathbf{S}(\mathbf{x}_i)$ can be expressed as

$$\mathbf{S}(\mathbf{x}_i) = [\mathbf{x}_i, \mathbf{x}_i^2, \mathbf{x}_i^3, (\mathbf{x}_i - \kappa_1)_+^3, \ldots, (\mathbf{x}_i - \kappa_5)_+^3].$$

The random vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ are taken from a multivariate normal distribution with zero mean and covariance matrix $\mathbf{I}_{\mathbf{d}\times\mathbf{d}}$, where $d = J + 3 = 8$. Then $\mathbf{y}_i$ from $\mathbf{y}_i = \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$ is generated for $i = 1, \ldots, n$. To examine the performance of the proposal method, the experiment is repeated 100 times each of sample size n = 100, 200, 500, and 1000.

To examine the performance in estimating $\Sigma_{\mathbf{x}}$, consider the following three criteria to evaluate the estimation accuracy of $\hat{\Sigma}_{\mathbf{x}}$, the estimate of $\Sigma_{\mathbf{x}}$:

$$\triangle_1 = \frac{1}{n}\sum_{i=1}^{n} \left[\mathrm{tr}\{\mathbf{H}(\mathbf{x}_i)\} - \log|\mathbf{H}(\mathbf{x}_i)|\right] - p, \tag{3.46}$$

$$\triangle_2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\{\mathbf{H}(\mathbf{x}_i) - \mathbf{I}\}^2, \tag{3.47}$$

$$\triangle_3 \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(\hat{\Sigma}_{\mathbf{x}_i} - \Sigma_{\mathbf{x}_i})^2 \tag{3.48}$$

where $\mathbf{H}(\mathbf{x}_i) = \Sigma_{\mathbf{x}_i}^{-1}\hat{\Sigma}_{\mathbf{x}_i}$, $|\mathbf{H}(\mathbf{x}_i)|$ is the determinant of matrix $\mathbf{H}(\mathbf{x}_i)$ and $\mathrm{tr}\{\mathbf{H}(\mathbf{x}_i)\}$ is its trace. Note that $\triangle_1$ and $\triangle_2$ are not the original Stein loss and the quadratic loss.

Table 3.1 summarize respectively the average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$, and $\triangle_3$ over 100 runs.

Table 3.1. The performance of the covariance regression model in estimating $\Sigma_{\mathbf{x}}$ for simulation example 1. The average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$ and $\triangle_3$ over 100 repetitions when sample size equals to 100, 200, 500, and 1000.

sample size		$\triangle_1$	$\triangle_2$	$\triangle_3$
100	average	0.2430	0.5470	1.6266
	stdev.	0.1219	0.4026	0.8894
	median	0.2065	0.4247	1.3833
200	average	0.1243	0.2697	0.8666
	stdev.	0.0664	0.1703	0.5215
	median	0.1050	0.2192	0.7524
500	average	0.0621	0.1296	0.4730
	stdev.	0.0278	0.0670	0.2332
	median	0.0591	0.1213	0.4487
1000	average	0.0445	0.0907	0.3976
	stdev.	0.0083	0.0194	0.1076
	median	0.0443	0.0879	0.3975

It can be seen from Table 3.1 that as the sample size n increases, the average, the standard deviation, and the median of $\triangle_1$, $\triangle_2$ and $\triangle_3$ would decrease, namely, the performance of the proposed method would be better.

Example 2

Now consider another simulation example for covariance model (3.1) with randomly generated parameter matrix $\mathbf{A}$ and $\mathbf{B}$. Suppose p and q are set the same as above. The generated parameter matrixes $\mathbf{A}$ and $\mathbf{B}$, satisfy 1) $\mathbf{A}$ is symmetric, each diagonal element of $\mathbf{A}$ is positive; 2) $\mathbf{B}$ has full rank, each element of $\mathbf{B}$ is between -1 and 1, and the

columns of **B** are orthogonal, which can be written as follows.

$$\mathbf{A} = \begin{bmatrix} 0.7864 & -0.9386 & -0.4500 & 0.2103 \\ -0.9386 & 1.6076 & 0.2483 & 0.6582 \\ -0.4500 & 0.2483 & 1.0292 & -0.2794 \\ 0.2103 & 0.6582 & -0.2794 & 2.0083 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.2649 & 0.7592 & 0.4719 \\ 0.4005 & 0.3779 & -0.2148 \\ 0.7400 & -0.1535 & -0.4961 \\ -0.4710 & 0.5072 & -0.6965 \end{bmatrix},$$

Similarly, the elements of explanatory samples **X** are generated from standard uniform distribution $U(0,1)$. The random error vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are taken from a multivariate normal population with zero mean and covariance matrix **A**. Five knots are selected at equal intervals over $[0,1]$. So the smoothing spline $\mathbf{S}(\mathbf{x_i})$ is of the same form as that in Example 1. The random vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ are taken from a multivariate normal distribution with zero mean and covariance matrix $\mathbf{I}_{d\times d}$ with $d = J+3 = 8$. The response variable $\mathbf{y}_i$ can then be generated from $\mathbf{y}_i = \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n$. To examine the performance of our proposal, the experiment is repeated 100 times each of sample size n = 100, 200, 500, and 1000.

The results of the average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$, and $\triangle_3$ over 100 repetitions are shown in Table 3.2.

It can be seen from Table 3.2 that as the sample size n increase, the average, the standard deviation, and the median of $\triangle_1$, $\triangle_2$ and $\triangle_3$ would decrease, that is, the performance of the proposed method would be better.

Example 3

Consider a simulation example for covariance regression model (3.37). Suppose p is 5, q is 4, and r is 3. First, the parameter matrixes **A**, **B**, and **C** are generated, such that 1) **A** is symmetric, each diagonal element of **A** is positive; 2) **B** has full rank, each element of **B** is between -1 and 1, and the columns of **B** are orthogonal; 3) **C** has full rank, each element of **C** is between -1 and 1, and the columns of **C** are orthogonal. **A**

Table 3.2. The performance of the covariance regression model in estimating $\Sigma_{\mathbf{x}}$ for simulation example 2. The average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$ and $\triangle_3$ over 100 repetitions when sample size equals to 100, 200, 500, and 1000.

sample size		$\triangle_1$	$\triangle_2$	$\triangle_3$
100	average	0.4607	6.3310	4.4998
	stdev.	0.6287	5.3844	3.7046
	median	0.2650	0.6064	3.2132
200	average	0.1695	0.8717	1.7059
	stdev.	0.1931	2.4070	1.2060
	median	0.1314	0.3487	1.4400
500	average	0.0685	0.3125	0.6556
	stdev.	0.0212	0.5616	0.2451
	median	0.0693	0.2111	0.6254
1000	average	0.0467	0.3347	0.3559
	stdev.	0.0159	0.4154	0.1505
	median	0.0429	0.1721	0.3210

B, and **C** can be written as follows.

$$\mathbf{A} = \begin{bmatrix} 2.4061 & 1.0373 & 0.8447 & 0.2015 & 0.3466 \\ 1.0373 & 2.6082 & 0.8604 & 0.2627 & 0.2260 \\ 0.8447 & 0.8604 & 2.3373 & 0.7075 & -0.5037 \\ 0.2015 & 0.2627 & 0.7075 & 0.4411 & -0.4116 \\ 0.3466 & 0.2260 & -0.5037 & -0.4116 & 0.6234 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} -0.2392 & 0.3585 & -0.7545 & -0.0776 \\ 0.5216 & -0.2955 & -0.1202 & 0.6855 \\ 0.7894 & 0.0954 & -0.2900 & -0.4899 \\ -0.1618 & -0.2555 & -0.5585 & 0.3203 \\ 0.1464 & 0.8425 & 0.1424 & 0.4261 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} -0.4914 & 0.7658 & -0.0029 \\ -0.3150 & -0.3145 & 0.7933 \\ 0.3255 & 0.0369 & 0.4117 \\ -0.5783 & -0.5486 & -0.3846 \\ -0.4679 & 0.1111 & 0.2306 \end{bmatrix},$$

where $\mathbf{A}$ is a 5×5 matrix, $\mathbf{B}$ is a 5×4 matrix, and $\mathbf{C}$ is a 5×3 matrix .

Then independently generate elements of $\mathbf{X}$ from standard uniform distribution $U(0,1)$ and elements of $\mathbf{Z}$ from Bernoulli distribution. The random error vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are taken from a multivariate normal population with zero mean and covariance matrix $\mathbf{A}$. Five knots are selected at equal intervals over $[0,1]$. So the smoothing spline $\mathbf{S}(\mathbf{x_i})$ is of the same form as that in Example 1. The random vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ are taken from a multivariate normal distribution with zero mean and covariance matrix $\mathbf{I}_{d\times d}$ with $d = J+3 = 8$. The random variables $\tau_1, \ldots, \tau_n$ are taken from a standard normal distribution. Then we generate $\mathbf{y}_i$ from $\mathbf{y}_i = \mathbf{B}\mathbf{S}(\mathbf{x}_i)\boldsymbol{\gamma}_i + \tau_i \mathbf{C}\mathbf{z}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n$. To examine the performance of the proposed method, the experiment is repeated 100 times each of sample size n = 100, 200, 500, and 1000.

To examine the performance in estimating $\Sigma_{\mathbf{x},\mathbf{z}}$, consider the following two criteria to evaluate the estimation accuracy of $\hat{\Sigma}_{\mathbf{x},\mathbf{z}}$, the estimate of $\Sigma_{\mathbf{x},\mathbf{z}}$:

$$\triangle_1 = \frac{1}{n}\sum_{i=1}^{n} \left[\mathrm{tr}\{\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i)\} - \log|\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i)|\right] - p, \tag{3.49}$$

$$\triangle_2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\{\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i) - \mathbf{I}\}^2, \tag{3.50}$$

$$\triangle_3 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(\hat{\Sigma}_{\mathbf{x}_i,\mathbf{z}_i} - \Sigma_{\mathbf{x}_i,\mathbf{z}_i})^2 \tag{3.51}$$

where $\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i) = \Sigma^{-1}_{\mathbf{x}_i,\mathbf{z}_i}\hat{\Sigma}_{\mathbf{x}_i,\mathbf{z}_i}$, $|\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i)|$ is the determinant of matrix $\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i)$, and additionally $\mathrm{tr}\{\mathbf{H}(\mathbf{x}_i,\mathbf{z}_i)\}$ is its trace. Note that $\triangle_1$ and $\triangle_2$ are are not the original Stein loss and the quadratic loss.

Table 3.3 summarize respectively the average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$, and $\triangle_3$ over 100 runs. It can be seen from Table 3.3 that as the sample size n increase, the average, the standard deviation and

the median of $\triangle_1$, $\triangle_2$, and $\triangle_3$ would decrease, that is, the performance of the proposed method would be better.

Table 3.3. The performance of the covariance regression model in estimating $\Sigma_{\mathbf{x}}$ for simulation example 3. The average ("average"), the standard deviation ("stdev.") and the median ("median") of $\triangle_1$, $\triangle_2$ and $\triangle_3$ over 100 repetitions when sample size equals to 100, 200, 500, and 1000.

sample size		$\triangle_1$	$\triangle_2$	$\triangle_3$
100	average	0.9550	3.9281	14.4240
	stdev.	0.4146	4.2178	7.4105
	median	0.8212	2.5336	12.4547
200	average	0.4705	1.6728	5.1726
	stdev.	0.0993	0.6354	1.7469
	median	0.4569	1.4974	5.1011
500	average	0.3273	1.3221	2.3855
	stdev.	0.0428	0.2955	0.6955
	median	0.3301	1.2961	2.2440
1000	average	0.2768	1.1691	1.5297
	stdev.	0.0262	0.1939	0.4105
	median	0.2746	1.1476	1.5177

3.4.2 Application to Boston Housing Data

In this section, we apply the proposed method on the Boston Housing data set for illustration . The data set reports the median value of owner-occupied homes in 506 U.S. census tracts in the Boston area in 1970, together with several variables which might help to explain the variation in housing value (see Harrison & Rubinfeld (1978)). For illustration purposes, we consider five social economics variables: CRIM(crime rate by town), TAX (full-value property-tax rate), PTRATIO (pupil-teacher ratio by town), MEDV (median value of owner-occupied homes), and NOX (nitric oxides concentration in parts per 10 million). For simplicity of notation, the response variables CRIM, TAX, PTRATIO, MEDV, and NOX are denoted by $y_1, y_2, \ldots, y_5$, respectively. Fan & Huang (2005) used $\sqrt{\text{LSTAT}}$ as the covariate x, where LSTAT denotes the percentage of lower status of the population. Here, the square-root transformation is employed since the distribution of LSTAT is examined to be asymmetric and therefore the resulting data have nearly symmetric distribution. Note that in order to be consistent with the model setting

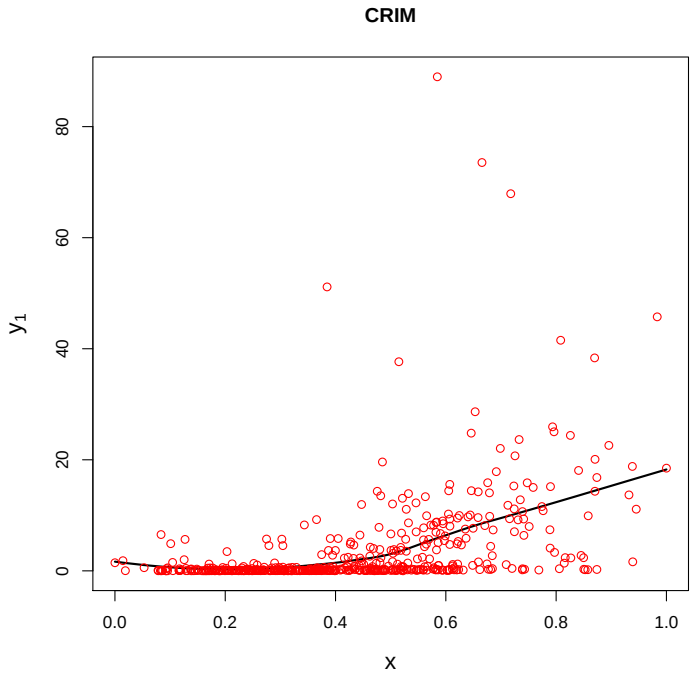

Figure 3.1. Estimated mean function for CRIM. The black line is local polynomial fits; the red dots indicate the observed value of the variable CRIM. These legends remain the same through Figure 3.5.

of the proposed model and reduce the possible bias, the regressor in the covariance function, x, is transformed so that it is uniformly distributed over $[0,1]$. To be specific, each observation vector is 5-dimensional with elements consisting of CRIM, TAX, PTRATIO, MEDV, and NOX respectively, which is denoted by $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i5})$. The explanatory variable x_i denotes the standardized square-root of LSTAT for the i-th observation. The objective of the study is to examine how the correlation structure of those y-variables varies as the percentage of lower status changes.

To evaluate the performance of the proposed nonparametric regression model, the mean function is needed to be estimated first. The the mean function can be subtracted from the corresponding data for response variables and the estimation of the covariance function can be proceeded. Precisely, if consider $\mathbf{y}_i | x_i \sim N(\boldsymbol{\mu}_{x_i}, \Sigma_{x_i})$, each element $\boldsymbol{\mu}_{x_i}$ is estimated by *loess* using the R-code, one for each of the response variable CRIM, TAX, PTRATIO, MEDV, and NOX, respectively. The plots of estimated regression functions $\boldsymbol{\mu}_{x_i}$ at those y-variables are shown in Figure 3.1 through Figure 3.5.

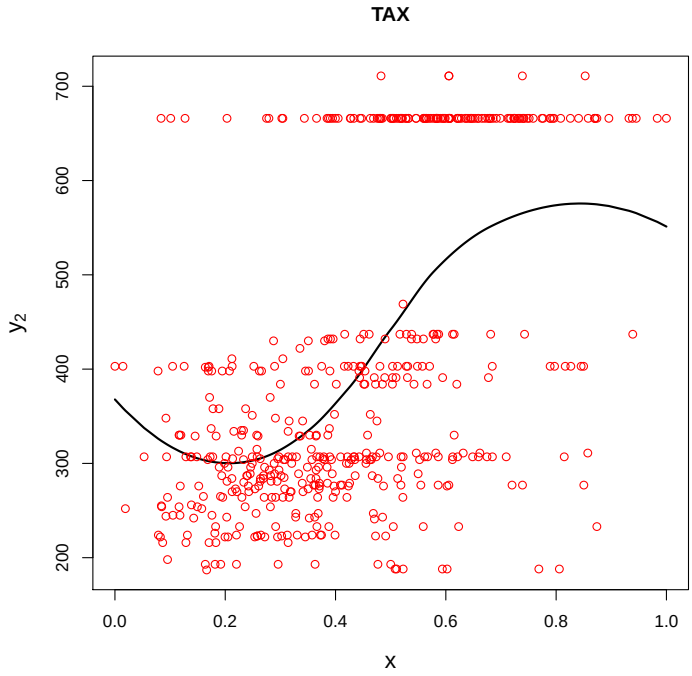

Figure 3.2. Estimated mean function for TAX.

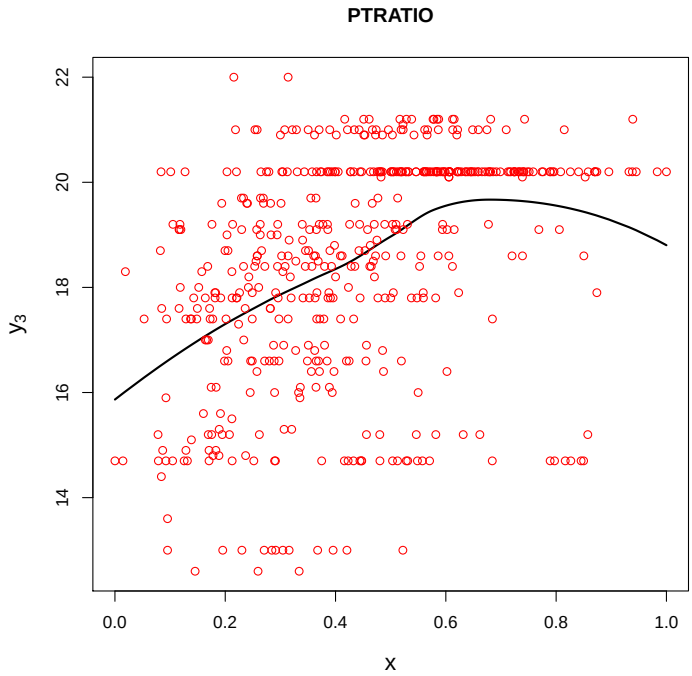

Figure 3.3. Estimated mean function for PTRATIO.

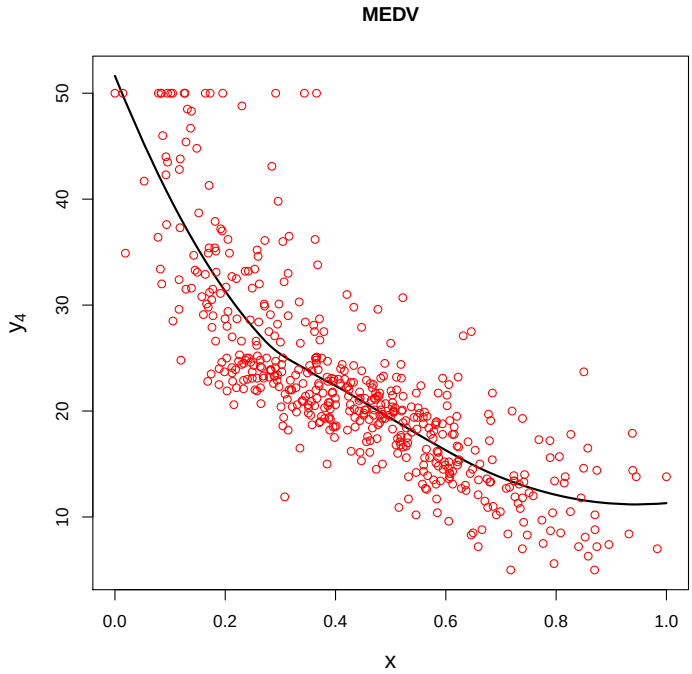

Figure 3.4. Estimated mean function for MEDV.

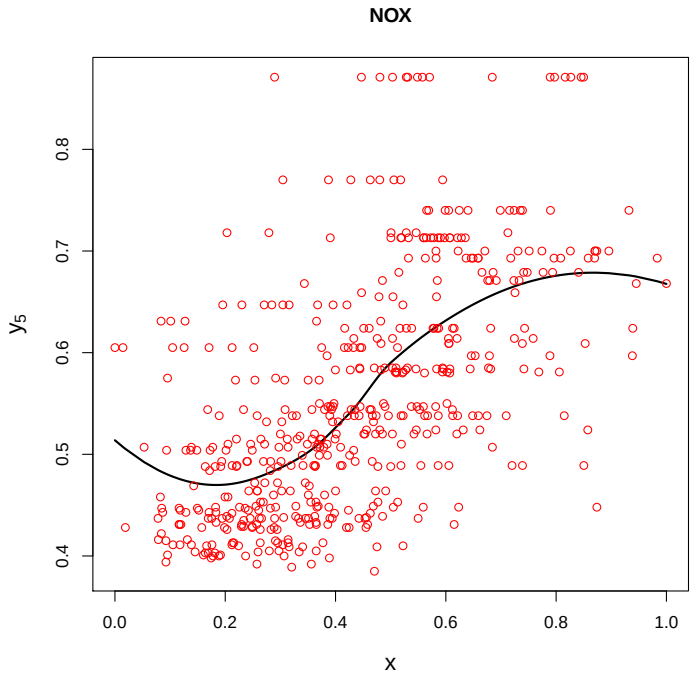

Figure 3.5. Estimated mean function for NOX.

Next, we obtain the conditional covariance matrix through the proposed nonparametric regression model. Five knots (i.e., $J = 5$), which denoted as $\kappa_1, \ldots, \kappa_5$, are selected at equal intervals over the range of x (i.e., [0,1]). Hence, the smoothing spline $\mathbf{S}(x_i)$ can be expressed as

$$\mathbf{S}(x_i) = [x_i, x_i^2, x_i^3, (x_i - \kappa_1)_+^3, \ldots, (x_i - \kappa_5)_+^3].$$

Then the pairwise nonparametric correlation coefficients are computed, which are denoted by $\hat{\rho}(y_1, y_2|x)$. Figure 3.6 through Figure 3.15 depict the estimation of conditional correlation coefficients. In these figures, the pointwise 95% confidence intervals $\hat{\rho}(y_1, y_2|x) \pm 1.96 \times \hat{\mathrm{SE}}\{\hat{\rho}(y_1, y_2|x)\}$, where the standard error estimate $\hat{\mathrm{SE}}\{\hat{\rho}_(y_1, y_2|x)\}$ was obtained based on 200 bootstrap experiments and 1.96 is approximately the 97.5 percentile point of the normal distribution. The legends are as follows: the solid black line is the estimated conditional correlation coefficients based on the proposed nonparametric covariance regression model; the dashed red line is the sample correlation coefficients based on the whole dataset; the dotted blue lines are the pointwise 95% confidence intervals based on the bootstrap experiments. For comparison purposes, the sample correlation coefficients are also presented in Table 3.4.

Table 3.4. Sample Correlation Coefficients.

	CRIM	TAX	PTRATIO	MEDV	NOX
CRIM	1.0000	0.4597	0.1709	-0.1780	0.2234
TAX	0.4597	1.0000	0.3192	-0.1414	0.4850
PTRATIO	0.1709	0.3192	1.0000	-0.3228	-0.0842
MEDV	-0.1780	-0.1414	-0.3228	1.0000	0.0469
NOX	0.2234	0.4850	-0.0842	0.0469	1.0000

Comparing the results in Table 3.4 and Figure 3.6 through Figure 3.15, a number of findings can be obtained. First, Table3.4 shows that sample correlation coefficients between the crime rate (CRIM) and the housing value (MEDV), the full-value property-tax rate (TAX) and the housing value (MEDV), the pupil-teacher ratio by town (PTRATIO) and the housing value (MEDV), and the pupil-teacher ratio by town (PTRATIO) and the nitric oxides concentration (NOX) , are negative. On the other hand, the correlation coefficients estimated by the proposed method present more details. To be specific, Figure 3.6 depicts that the correlation coefficient between the crime rate (CRIM) and the full-value property-tax rate (TAX) has a curved trend as lower status increases. When the lower status is small, CRIM and TAX has high correlation; as the lower status increases

this correlation coefficient decreases and it reaches its minimum when x is close to 0.5; then the correlation coefficient goes up as x increases. Analogously, the correlation coefficients between the crime rate (CRIM) and the pupil-teacher ratio by town (PTRATIO) (see Figure 3.7), the crime rate (CRIM), and the nitric oxides concentration (NOX) (see Figure 3.9) have similar trends as that between CRIM and TAX. Figure 3.8 shows that the correlation coefficient between the crime rate (CRIM) and the housing value (MEDV) has a decreasing trend as lower status increases. It goes from weak positive correlation from moderate negative correlation as the lower status grows. Furthermore, the correlation between the tax rate (TAX) and the housing value (MEDV) (see Figure 3.11) drops as the lower status increases. Similar trend can be observed from the housing value (MEDV) and the nitric oxides concentration (NOX) (see Figure 3.15). It is noted that these findings are not available from the simple sample correlation coefficient matrix presented in Table 3.4 .

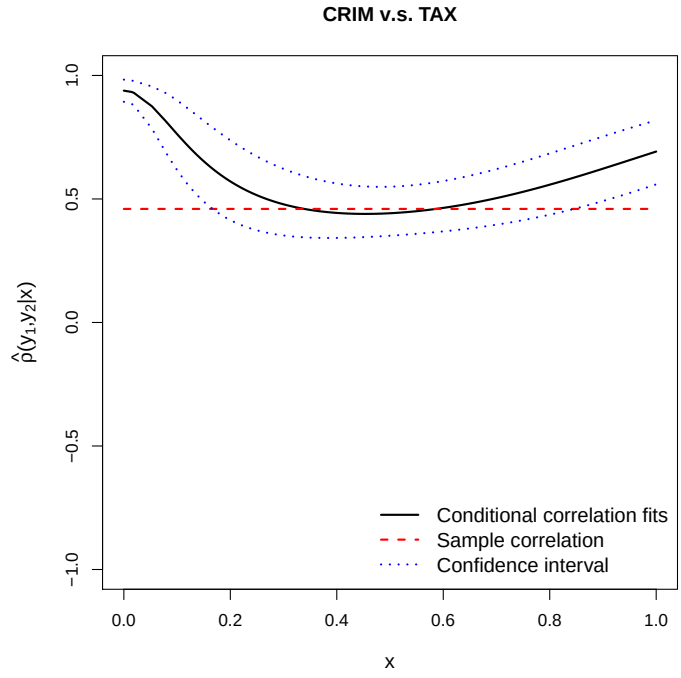

Figure 3.6. Estimated correlation coefficients of CRIM and TAX for Boston Housing Data. The solid black line is the estimated conditional correlation coefficients based on the proposed nonparametric covariance regression model; the dashed red line is the sample correlation coefficients based on the whole dataset; the dotted blue lines are the pointwise 95% confidence intervals based on the bootstrap experiments. These legends remain the same through Figure 3.15.

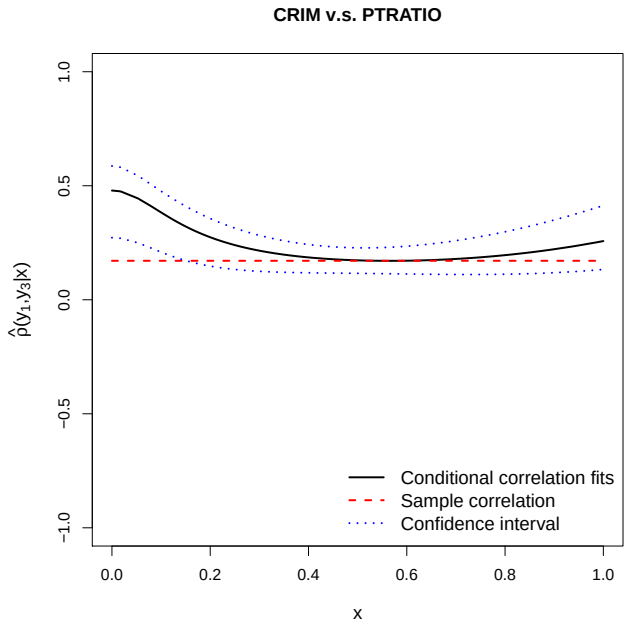

Figure 3.7. Estimated correlation coefficients of CRIM and PTRATIO for Boston Housing Data.

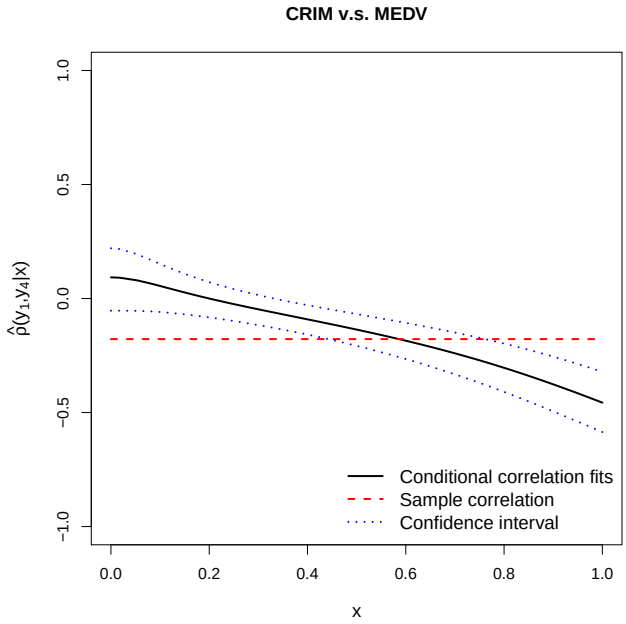

Figure 3.8. Estimated correlation coefficients of CRIM and MEDV for Boston Housing Data.

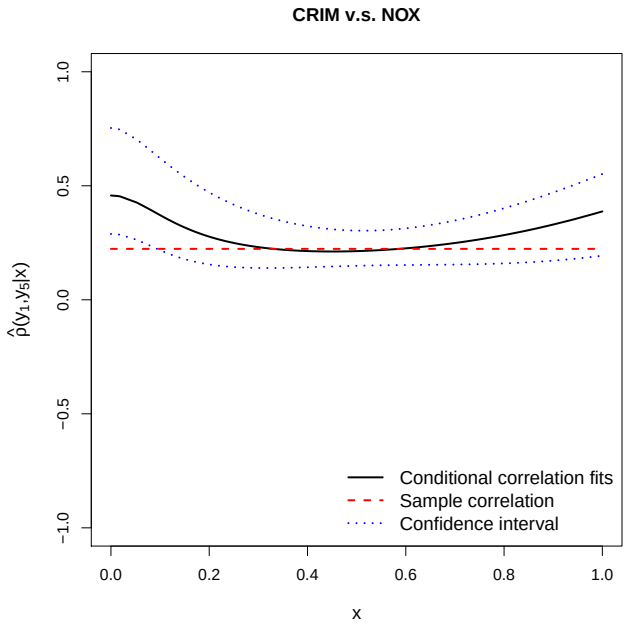

Figure 3.9. Estimated correlation coefficients of CRIM and NOX for Boston Housing Data.

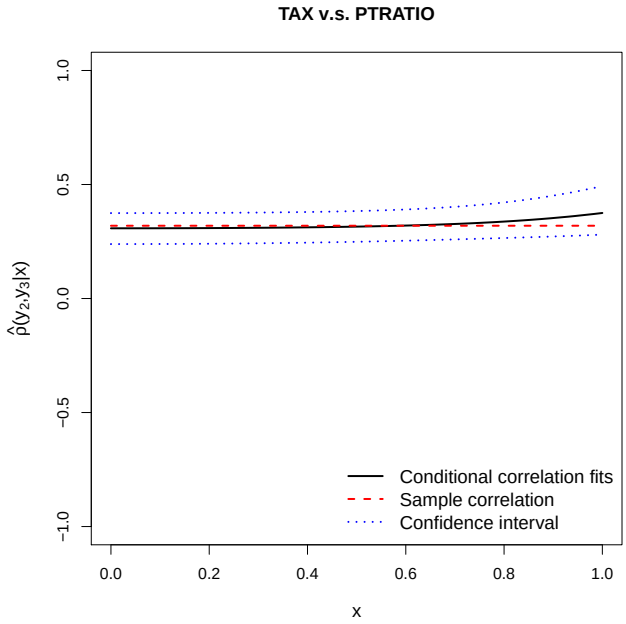

Figure 3.10. Estimated correlation coefficients of TAX and PTRATIO for Boston Housing Data.

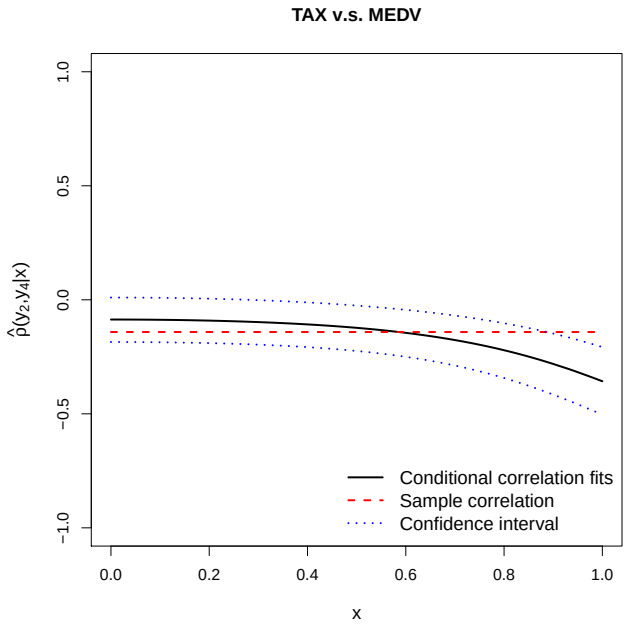

Figure 3.11. Estimated correlation coefficients of TAX and MEDV for Boston Housing Data.

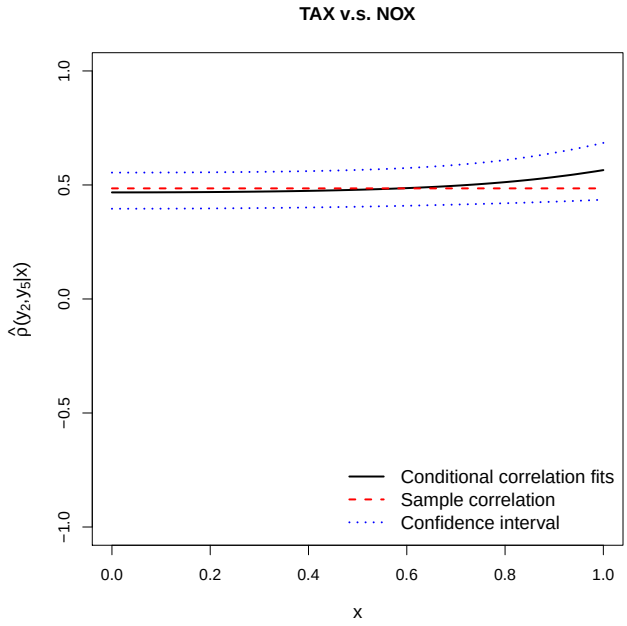

Figure 3.12. Estimated correlation coefficients of TAX and NOX for Boston Housing Data.

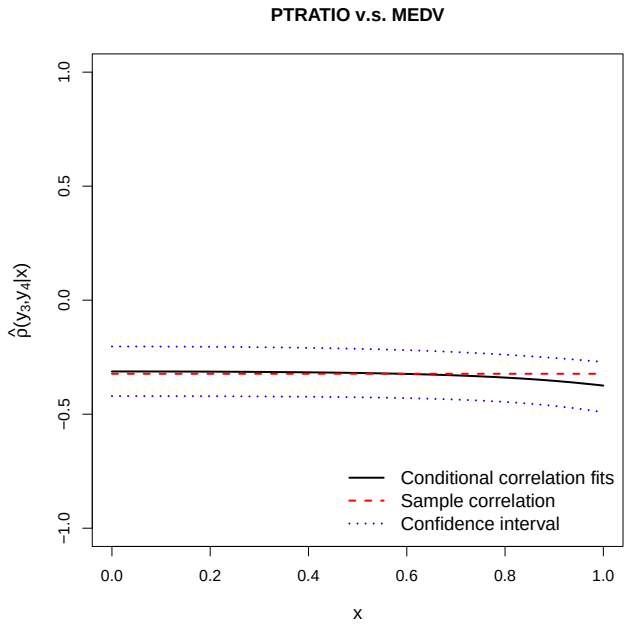

Figure 3.13. Estimated correlation coefficients of PTRATIO and MEDV for Boston Housing Data.

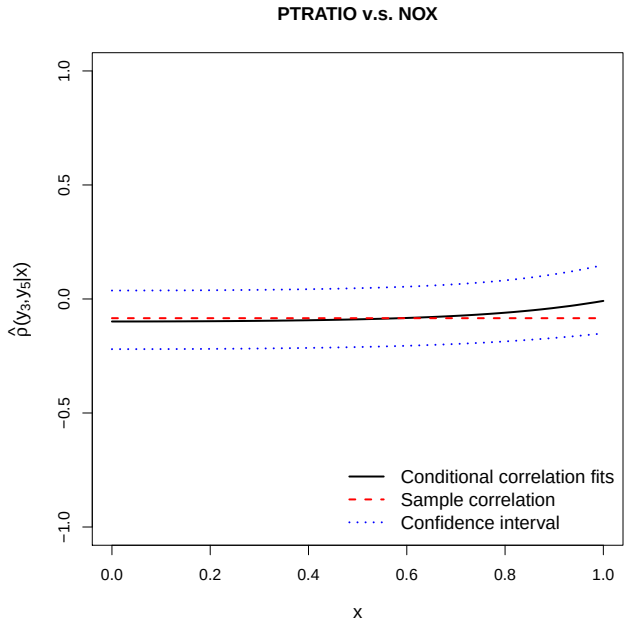

Figure 3.14. Estimated correlation coefficients of PTRARIO and NOX for Boston Housing Data.

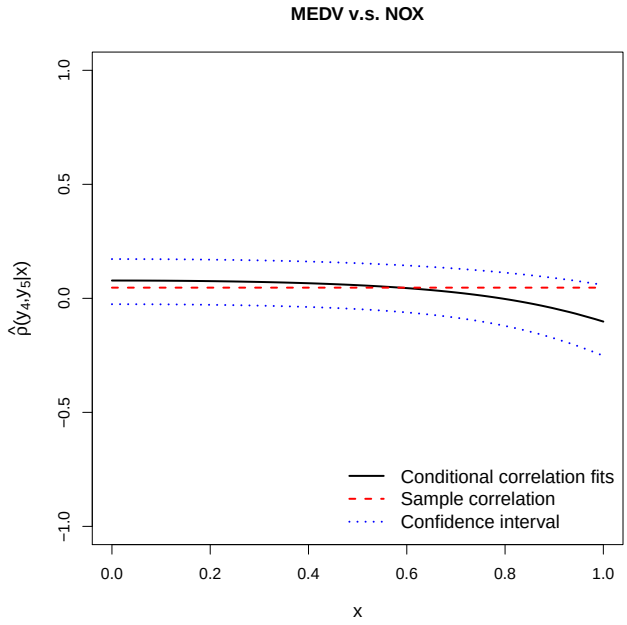

Figure 3.15. Estimated correlation coefficients of MEDV and NOX for Boston Housing Data.

Chapter 4

Functional Estimation of Conditional Covariance

Chapter 3 considers the situation in which the covariates are low dimensional. However, it becomes less useful in situations where the covariates are high-dimensional. In this chapter, estimating the conditional covariance matrix through a modified Cholesky decomposition is proposed. The modified Cholesky decomposition procedure associates each local covariance matrix with a unique unit lower triangular and a unique diagonal matrix. The entries of the lower triangular matrix and the diagonal matrix have statistical interpretation as regression coefficients and prediction variances when regressing each term on its predecessors. A class of partially linear models are used to estimate those regression coefficients and kernel estimators are developed to estimate the nonparametric variance functions. The asymptotic properties of the proposed procedure are studied. Comprehensive simulation studies are conducted to examine the finite sample performance of the proposed procedures. A real data example is used to illustrate the proposed methodology.

4.1 Estimation of Covariance Function Assuming Partially Linear Models

4.1.1 The Modified Cholesky Decomposition

Let $\mathbf{y} = (Y_1, \ldots, Y_m)^T$ be the response vector, $\mathbf{x} = (X_1, \ldots, X_d)^T$ and U be the associated covariates, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^T$ be the error vector with $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{x}, U) = \mathbf{0}$. Then in regression analysis, the responses usually can be decomposed into two uncorrelated parts as follows:

$$\mathbf{y} \quad = \quad \mathbb{E}(\mathbf{y}|\mathbf{x}, U) + \boldsymbol{\varepsilon}. \tag{4.1}$$

In the subsection, the estimation of the conditional mean function will be discussed. In order to illustrate the rationale of the proposed method, the conditional mean is assumed to be known for now. The primary interest in this chapter is to estimate the conditional covariance matrix $\boldsymbol{\Sigma}(\mathbf{x}, U) = \mathrm{Cov}(\mathbf{y}|\mathbf{x}, U) = \mathrm{Cov}(\boldsymbol{\varepsilon}|\mathbf{x}, U)$. The idea of Cholesky decomposition through associating the conditional covariance matrix with a unique unit lower triangular and a unique diagonal matrix will be adopted. To be more specific, since $\boldsymbol{\Sigma}(\mathbf{x}, U)$ is symmetric, it has the following modified Cholesky decomposition structure:

$$\mathbf{L}(\mathbf{x}, U)\boldsymbol{\Sigma}(\mathbf{x}, U)\mathbf{L}^T(\mathbf{x}, U) = \mathbf{D}(\mathbf{x}, U), \tag{4.2}$$

where $\mathbf{L}(\mathbf{x}, U)$ is a lower triangular matrix containing ones on its diagonal and elements $-\phi_{kj}(\mathbf{x}, U)$ in the (k, j)-th position for $1 \leq j < k \leq m$, and $\mathbf{D}(\mathbf{x}, U)$ is a diagonal matrix of diagonals $\sigma_1^2(\mathbf{x}, U), \ldots, \sigma_m^2(\mathbf{x}, U)$.

The modified Cholesky decomposition (4.2) can be re-written as

$$\boldsymbol{\Sigma}(\mathbf{x}, U) = \{\mathbf{L}(\mathbf{x}, U)\}^{-1}\mathbf{D}(\mathbf{x}, U)\{\mathbf{L}^T(\mathbf{x}, U)\}^{-1}. \tag{4.3}$$

Therefore, in consideration of estimating the covariance matrix using the Cholesky decomposition (4.2), it is sufficient to estimate the entries of the lower triangular matrix $\mathbf{L}(\mathbf{x}, U)$ and the diagonal matrix $\mathbf{D}(\mathbf{x}, U)$, particularly, $-\phi_{kj}(\mathbf{x}, U), 1 \leq j < k \leq m$, and $\sigma_k^2(\mathbf{x}, U), k = 1, \ldots, m$. Based on the form of lower triangular matrix $\mathbf{L}(\mathbf{x}, U)$, the Cholesky decomposition (4.2) can be re-formulated through regression modeling. To be

concrete, denote

$$\mathbf{e} = \mathbf{L}(\mathbf{x}, U)\boldsymbol{\varepsilon}, \tag{4.4}$$

i.e., $e_1 = \varepsilon_1$ and $e_k = \varepsilon_k - \sum_{j=1}^{k-1} \phi_{kj}(\mathbf{x}, U)\varepsilon_j$ for $k = 2, \cdots, m$, where the regression coefficients ϕ_{kj}'s are unconstrained and the variances $\mathrm{Var}(e_k|\mathbf{x}, U) = \sigma_k^2(\mathbf{x}, U)$ are non-negative. Then,

$$\varepsilon_1 = e_1, \quad \varepsilon_k = \sum_{j=1}^{k-1} \phi_{kj}(\mathbf{x}, U)\varepsilon_j + e_k, \;\; k \geq 2. \tag{4.5}$$

Evidently, the prediction errors are uncorrelated since $\mathrm{Cov}(\mathbf{e}|\mathbf{x}, U) = \mathbf{D}(\mathbf{x}, U)$ is a diagonal matrix. Therefore, the elements $\phi_{kj}(\mathbf{x}, U)$ in the lower triangular matrix $\mathbf{L}(\mathbf{x}, U)$ and the diagonals $\sigma_k^2(\mathbf{x}, U)$ of $\mathbf{D}(\mathbf{x}, U)$ have statistical interpretation as the regression coefficients and prediction variances corresponding to regressing each error term ε_k on its predecessor $\varepsilon_j, j = 1, \ldots, k$ for $k = 2, \ldots, m$.

First we focus on estimating the coefficients $\phi_{kj}(\mathbf{x}, U)$. There has been focused research in the literature for local smoothing techniques when the dimension $(d+1)$ of the covariates $(\mathbf{x}, U)$ is relatively small and the sample size n is sufficiently large, see for example, the local polynomial regression by Fan & Gijbels (1996). However, in modern data sets, the dimension of covariates is usually large, which requires new techniques to estimate $\phi_{kj}(\mathbf{x}, U)$. In this proposal, we consider model the regression of $\phi(\mathbf{x}, U)$ through partially linear models. To be specific, assume

$$\phi_{kj}(\mathbf{x}, U) = \beta_{kj}(U) + \boldsymbol{\gamma}_{kj}^T \mathbf{x} \tag{4.6}$$

for $k = 2, \ldots, m, \;\; j = 1, \ldots, k-1$, where $\beta_{kj}(\cdot)$'s are unknown smoothing functions and $\boldsymbol{\gamma}_{kj}$'s are $d \times 1$ vectors of unknown parameters. Then the error term (4.5) can be re-written as that for $k = 2, \ldots, m$,

$$\begin{aligned} \varepsilon_k &= \sum_{j=1}^{k-1} \phi_{kj}(\mathbf{x}, U)\varepsilon_j + e_k \\ &= \sum_{j=1}^{k-1} \left\{ \beta_{kj}(U) + \boldsymbol{\gamma}_{kj}^T \mathbf{x} \right\} \varepsilon_j + e_k \end{aligned}$$

$$= \boldsymbol{\varepsilon}_{(k-1)}^T \boldsymbol{\beta}_{k,(k-1)}(U) + (\boldsymbol{\varepsilon}_{(k-1)} \otimes \mathbf{x})^T \boldsymbol{\gamma}_k + e_k, \tag{4.7}$$

where the symbol $\otimes$ denotes the Kronecker product,

$$\boldsymbol{\varepsilon}_{(k-1)} = (\varepsilon_1, \ldots, \varepsilon_{k-1})^T, \boldsymbol{\beta}_{k,(k-1)}(\cdot) = \big(\beta_{k1}(\cdot), \ldots, \beta_{k,k-1}(\cdot)\big)^T, \boldsymbol{\gamma}_{kj} = (\gamma_{kj1}, \ldots, \gamma_{kjd})^T,$$

and $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}^T, \ldots, \boldsymbol{\gamma}_{k,k-1}^T)^T$ stacks the vectors $\boldsymbol{\gamma}_{kj}$ from $j = 1$ through $k-1$ into a long column vector. The expression (4.7) is a semiparametric varying-coefficient partially linear model, which has been well studied in the literature, see for example Fan & Huang (2005), Fan et al. (2007), etc.. Profile least square techniques can be used for estimation of the partially linear varying-coefficient models. The details of estimation procedures will be discussed in the subsection. With efficient estimators of $\beta_{kj}(U)$ and $\boldsymbol{\gamma}_{kj}$'s, the elements $\phi_{kj}(\mathbf{x}, U)$ can be estimated efficiently. Consequently, the estimation of $\mathbf{L}(\mathbf{x}, U)$ can be obtained.

Next we consider estimating the elements $\sigma_k^2(\mathbf{x}, U)$ in the diagonal matrix $\mathbf{D}(\mathbf{x}, U)$. From previous discussion, it is noted that $\sigma_k^2(\mathbf{x}, U) = \mathrm{Var}(e_k|\mathbf{x}, U)$, for $k = 1, \ldots, m$. In view of simplicity, we assume that $\sigma_k^2(\mathbf{x}, U) = \sigma_k^2(U)$ throughout the present context. Accordingly, $\sigma_k^2(U)$ can be directly estimated by using standard nonparametric smoothing techniques. A natural estimator for $\sigma_k^2(U)$ is the kernel estimator, which ensures that the estimated covariance function is positive definite locally. The kernel estimate for $\sigma_k^2(U)$ can be constructed as follows:

$$\hat{\sigma}_k^2(U) = \frac{\sum_{i=1}^n \hat{e}_{ik}^2 K_{h_2}(U_i - U)}{\sum_{i=1}^n K_{h_2}(U_i - U)}, \; k = 1, \ldots, m, \tag{4.8}$$

where $K_{h_2}(\cdot) = K(\cdot/h_2)/h_2$, $K(\cdot)$ is a kernel density function, and h_2 is a smoothing parameter.

4.1.2 Sample Estimation of Conditional Covariance Matrix

There are many approaches to estimate the unknown parameters $\{\beta_{kj}, k = 2, \ldots, m, j = 1, \ldots, k-1\}$ and the varying coefficient functions $\{\gamma_{kjl}, k = 2, \ldots, m, j = 1, \ldots, k-1, l = 1, \ldots, d\}$. Profile least square is a useful approach for estimation of the partially linear varying-coefficient model.

Suppose that we have a random sample of size n, $\{(U_i, \mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$, from

the population $(U, \mathbf{x}, \mathbf{y})$, where $\mathbf{x}_i = (X_{i1}, \ldots, X_{id})^T$ and $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{im})^T$. First, we assume that the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{x} = \mathbf{x}_i, U = U_i)$ for $i = 1, \ldots, n$, are observed. Accordingly, the error ε_i's are assumed to be known as a priori. In this subsection, we will focus on the sample estimation for $\mathbf{L}(\mathbf{x}, U)$ and $\mathbf{D}(\mathbf{x}, U)$ with true errors ε_i's. We will also discuss the replacement of true errors with corresponding residuals for the sample estimation. Profile least square techniques will be used for estimation of the partially linear varying-coefficient models. Practically, we can give a closed form for the profile least-squares estimator of $(\boldsymbol{\beta}_{k,(k-1)}, \boldsymbol{\gamma}_k)$ based on the given samples for each $k = 2, \ldots, m$.

Denote $\mathbf{v}_k = \boldsymbol{\varepsilon}_{(k-1)} \otimes \mathbf{x}$, a column vector of $d(k-1)$ components. Then (4.7) can be re-written as

$$\varepsilon_k = \boldsymbol{\varepsilon}_{(k-1)}^T \boldsymbol{\beta}_{k,(k-1)}(U) + \mathbf{v}_k^T \boldsymbol{\gamma}_k + e_k. \tag{4.9}$$

For a given $\boldsymbol{\gamma}_k$, let $\varepsilon_k^* = \varepsilon_k - \mathbf{v}_k^T \boldsymbol{\gamma}_k$. Thus

$$\varepsilon_k^* = \boldsymbol{\varepsilon}_{(k-1)}^T \boldsymbol{\beta}_{k,(k-1)}(U) + e_k. \tag{4.10}$$

Recall that $e_1, \ldots, e_m$ are uncorrelated in the modified Cholesky decomposition. Thus (4.10) is a varying coefficient model, assuming that ε_k^* and $\boldsymbol{\varepsilon}_{(k-1)}$ are known.

Let $\varepsilon_{i,k}$, $\mathbf{v}_{i,k}$, and $e_{i,k}$ be the samples of ε_k, $\mathbf{v}_k$, and e_k corresponding to the i-th sample, respectively, and let $m_{i,k} = \boldsymbol{\varepsilon}_{i,(k-1)}^T \boldsymbol{\beta}_{k,(k-1)}(U_i)$ for $i = 1, \ldots, n$. Put the samples together in matrix form, we have the sample model of (4.9) in matrix form as

$$\boldsymbol{\varepsilon}_k - \mathbf{V}_k \boldsymbol{\gamma}_k = \mathbf{m}_k + \mathbf{e}_k, \tag{4.11}$$

where

$$\boldsymbol{\varepsilon}_k = (\varepsilon_{1k}, \ldots, \varepsilon_{nk})^T,\ \mathbf{V}_k = (\mathbf{v}_{1k}, \ldots, \mathbf{v}_{nk})^T,\ \mathbf{e}_k = (e_{1k}, \ldots, e_{nk})^T,\ \mathbf{m}_k = (m_{1k}, \ldots, m_{nk})^T.$$

It is known that the local linear regression results in a linear estimate for $\boldsymbol{\beta}_{k,(k-1)}(\cdot)$. Hence, the estimate of $\boldsymbol{\beta}_{k,(k-1)}(\cdot)$ is linear in $\boldsymbol{\varepsilon}_k - \mathbf{V}_k \boldsymbol{\gamma}_k$, and thus the estimate of $\mathbf{m}_k$ is of the form $\mathbf{S}_h(\boldsymbol{\varepsilon}_k - \mathbf{V}_k \boldsymbol{\gamma}_k)$. The matrix $\mathbf{S}_h$ is known as a smoothing matrix of the local

linear smoother with bandwidth h. In addition, the profile least square is as follows:

$$\begin{aligned}\ell_k(\boldsymbol{\beta}_{k,(k-1)},\boldsymbol{\gamma}_k) &= \|\boldsymbol{\varepsilon}_k-\mathbf{V}_k\boldsymbol{\gamma}_k-\mathbf{S}_h(\boldsymbol{\varepsilon}_k-\mathbf{V}_k\boldsymbol{\gamma}_k)\|^2\\ &= \|(\mathbf{I}-\mathbf{S}_h)(\boldsymbol{\varepsilon}_k-\mathbf{V}_k\boldsymbol{\gamma}_k)\|^2. \end{aligned} \tag{4.12}$$

We can employ local linear regression techniques to estimate $\boldsymbol{\beta}_{k,(k-1)}(\cdot)$ for each $k \geq 2$. To be specific, for any U in the neighborhood of a given U_0, we can locally approximate $\beta_{kj}(U)$ by a linear function as follows:

$$\beta_{kj}(U) \approx a_{kj} + b_{kj}(U-U_0). \tag{4.13}$$

The local linear regression is to find the coefficients

$$\mathbf{a}_k = (a_{k1},\ldots,a_{k,k-1})^T, \quad \mathbf{b}_k = (b_{k1},\ldots,b_{k,k-1})^T$$

approximately by minimizing the following least squares functions

$$\hat{\ell}_k(\mathbf{a}_k,\mathbf{b}_k) = \sum_{i=1}^{n}\left[\varepsilon_{ik}^* - \sum_{j=1}^{k-1}\left\{a_{kj}+b_{kj}(U_i-U_0)\right\}\varepsilon_{ij}\right]^2 K_h(U_i-U_0), \tag{4.14}$$

where $\varepsilon_{i,k}^* = \varepsilon_{i,k} - \mathbf{v}_{i,k}^T\boldsymbol{\gamma}_k$ and $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function with bandwidth h. The optimal solution of $\min_{\mathbf{a}_k,\mathbf{b}_k}\hat{\ell}_k(\mathbf{a}_k,\mathbf{b}_k)$ is given by

$$(\hat{\mathbf{a}}_k^T,\hat{\mathbf{b}}_k^T)^T = \mathbf{H}_h(U_0)(\boldsymbol{\varepsilon}_k-\mathbf{V}_k\boldsymbol{\gamma}_k), \tag{4.15}$$

where

$$\begin{aligned}\mathbf{H}_h(U_0) &= \left(\mathbf{D}(U_0)^T\mathbf{K}(U_0)\mathbf{D}(U_0)\right)^{-1}\mathbf{D}(U_0)^T\mathbf{K}(U_0),\\ \mathbf{K}(U_0) &= \mathrm{diag}\{K_h(U_1-U_0),\ldots,K_h(U_n-U_0)\},\\ \mathbf{D}(U_0) &= \left[\mathbf{E}_k,\ \mathrm{diag}\{(U_1-U_0)/h,\ldots,(U_n-U_0)/h\}\mathbf{E}_k\right],\end{aligned}$$

with $\mathbf{E}_k = [\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_{k-1}]$. The optimal solution yields a local linear estimate for $\boldsymbol{\beta}_{k,(k-1)}(U)$ near U_0:

$$\hat{\boldsymbol{\beta}}_{k,(k-1)}(U) \approx \hat{\mathbf{a}}_k + \hat{\mathbf{b}}_k(U-U_0).$$

The local linear regression gives an approximate $\hat{\mathbf{m}}_k = \mathbf{S}_h(\boldsymbol{\varepsilon}_k - \mathbf{V}_k\boldsymbol{\gamma}_k)$ to $\mathbf{m}_k$ with

$$\mathbf{S}_h = \begin{bmatrix} (\boldsymbol{\varepsilon}_{1(k-1)}, \mathbf{0})\mathbf{H}_h(U_1) \\ \vdots \\ (\boldsymbol{\varepsilon}_{n(k-1)}, \mathbf{0})\mathbf{H}_h(U_n) \end{bmatrix}, \tag{4.16}$$

and the regression error

$$\hat{\mathbf{e}}_k = (\mathbf{I} - \mathbf{S}_h)(\boldsymbol{\varepsilon}_k - \mathbf{V}_k\boldsymbol{\gamma}_k). \tag{4.17}$$

The parameter vectors $\boldsymbol{\gamma}_k$, $k = 2, \ldots, m$, can be regressed recursively as follows. Taking the sum of the norm squares in (4.17) , we have that total error

$$\mathbf{e}_{\text{total}}(\boldsymbol{\gamma}_k) = \left\|(\mathbf{I} - \mathbf{S}_h)(\boldsymbol{\varepsilon}_k - \mathbf{V}_k\boldsymbol{\gamma}_k)\right\|_2^2.$$

The linear regression minimizing the error $\mathbf{e}_{\text{total}}(\boldsymbol{\gamma}_k)$ gives the following approximate to $\boldsymbol{\gamma}_k$:

$$\hat{\boldsymbol{\gamma}}_k = \left\{\mathbf{V}_k^T(\mathbf{I} - \mathbf{S}_h)^T(\mathbf{I} - \mathbf{S}_h)\mathbf{V}_k\right\}^{-1}\left\{\mathbf{V}_k^T(\mathbf{I} - \mathbf{S}_h)^T(\mathbf{I} - \mathbf{S}_h)\boldsymbol{\varepsilon}_k\right\}. \tag{4.18}$$

Substituting the above estimates into (4.15), we obtain the estimates of $\boldsymbol{\beta}_{k,(k-1)}(U)$ and $\dot{\boldsymbol{\beta}}_{k,(k-1)}(U)$,

$$\hat{\boldsymbol{\beta}}_{k,(k-1)}(U) = [\mathbf{I}, \mathbf{0}]\mathbf{H}_k(U)(\boldsymbol{\varepsilon}_k - \mathbf{V}_k\hat{\boldsymbol{\gamma}}_k), \tag{4.19}$$

$$\hat{\dot{\boldsymbol{\beta}}}_{k,(k-1)}(U) = [\mathbf{0}, \mathbf{I}]\mathbf{H}_k(U)(\boldsymbol{\varepsilon}_k - \mathbf{V}_k\hat{\boldsymbol{\gamma}}_k). \tag{4.20}$$

Therefore, the regression coefficient $\phi_{kj}(\mathbf{x}, U)$ can be estimated with

$$\hat{\phi}_{kj}(\mathbf{x}, U) = \hat{\beta}_{kj}(U) + \hat{\boldsymbol{\gamma}}_{kj}^T\mathbf{x} \tag{4.21}$$

for $j = 1, \ldots, k-1$. Consequently, the estimation of $\mathbf{L}(\mathbf{x}, U)$ can be obtained by replacing the above procedure for $k = 2, \ldots, m$.

Thereafter we will investigate the estimate for $\sigma_k^2(U)$ in the diagonal matrix $\mathbf{D}(\mathbf{x}, U)$. The standard kernel smoothing techniques can be applied to estimate $\sigma_k^2(U)$ as $\beta_{kj}(U)$'s

and γ_{kj}'s are obtained. To be specific, we define the residuals by

$$\hat{e}_{i1} = \varepsilon_{i1},$$
$$\hat{e}_{ik} = \varepsilon_{ik} - \sum_{j=1}^{k-1} \left\{ \hat{\beta}_{kj}(U_i) + \hat{\boldsymbol{\gamma}}_{kj}^T \mathbf{x}_i \right\} \varepsilon_{ij},$$

for $i = 1,\ldots,n$ and $k = 2,\ldots,m$. Therefore, the kernel estimator for $\sigma_k^2(U)$ can be directly constructed by

$$\hat{\sigma}_k^2(U) = \frac{\sum_{i=1}^n \hat{e}_{ik}^2 K_{h_2}(U_i - U)}{\sum_{i=1}^n K_{h_2}(U_i - U)},\ k = 1,\ldots,m, \tag{4.22}$$

where $K_{h_2}(\cdot) = K(\cdot/h_2)/h_2$, $K(\cdot)$ is a kernel density function, and h_2 is a smoothing parameter known as a bandwidth. Subsequently, a proper estimate of $\mathbf{D}(\mathbf{x},U)$ is a diagonal matrix $\hat{\mathbf{D}}(\mathbf{x},U)$ with its (k,k)-th element being $\hat{\sigma}_k^2(U)$, for $k = 1,\ldots,m$.

With the obtained estimators $\hat{\mathbf{L}}(\mathbf{x},U)$ and $\hat{\mathbf{D}}(\mathbf{x},U)$, we can estimate $\boldsymbol{\Sigma}(\mathbf{x},U)$ by its Cholesky decomposition $\hat{\boldsymbol{\Sigma}}(\mathbf{x},U) = \{\hat{\mathbf{L}}(\mathbf{x},U)\}^{-1}\hat{\mathbf{D}}(\mathbf{x},U)\{\hat{\mathbf{L}}^T(\mathbf{x},U)\}^{-1}$.

4.1.3 Estimation of Conditional Mean

In previous subsections, the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$ is assumed to be directly observable. This facilitates to illustrate the motivation of our proposed method. However, in practice, the conditional mean function is usually unobservable. Consequently, it is not trivial to provide a consistent estimator for $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$, especially when the dimension of the covariates $(\mathbf{x},U)$ $(d+1)$ is fairly large compared with the sample size n. When the dimension $(d+1)$ is relatively small, standard nonparametric smoothing techniques such as the local polynomial regression (Fan & Gijbels (1996)) can be used to efficiently estimate the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$. In this subsection, we will mainly focus on estimating the conditional mean functions when the dimension $(d+1)$ is large. The varying coefficient models of the form

$$Y_j = \alpha_{j0}(U) + \mathbf{x}^T \boldsymbol{\alpha}_j(U) + \varepsilon_j,\ \text{for } j = 1,\ldots,m, \tag{4.23}$$

are proposed to model the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$, where $\alpha_{j0}(U)$ and $\boldsymbol{\alpha}_j(U) = \{\alpha_{j1}(U),\ldots,\alpha_{jd}(U)\}^T$ are smoothing functions of U. This varying coefficient model

is studied by Fan & Zhang (2000) in the context of longitudinal data and by Hastie & Tibshirani (1993) for the case of independent and identical distributed observations. This varying coefficient model maintains the modeling flexibility and simultaneously avoids the curse of dimensionality of covariates. In addition, the varying-coefficients can be delicately interpreted in many applications.

The unknown smoothing functions $\alpha_{j0}(U)$ and $\boldsymbol{\alpha}_j(U) = \{\alpha_{j1}(U), \ldots, \alpha_{jd}(U)\}^T$ can be easily estimated by using any linear smoother. Here we employ local linear approximation (Fan & Gijbels (1996)). To be specific, for any U_i in the neighborhood of a given U, we can locally approximate $\alpha_{il}(U)$ by following form Taylor's expansion,

$$\alpha_{jl}(U) \approx \alpha_{jl}(U_i) + \alpha'_{jl}(U_i)(U_i - U) \equiv c_{jl} + d_{jl}(U_i - U), \tag{4.24}$$

for $j = 1, \ldots, m$ and $l = 0, \ldots, d$. Note that since the data are localized in covariate U, the covariance structure does not greatly affect the local linear estimator Fan et al. (2007). Therefore, we can estimate $\{\alpha_{jl}(U), l = 0, \ldots, d\}$ for each j by minimizing the following local least squares function

$$\hat{\ell}_j(\mathbf{c}_j, \mathbf{d}_j) = \sum_{i=1}^{n} \Big[Y_{ij} - \sum_{l=0}^{d} \big\{ c_{jl} + d_{jl}(U_i - U) \big\} X_{il} \Big]^2 K_{h_1}(U_i - U), \tag{4.25}$$

where $X_{i0} \equiv 1$, $\mathbf{c}_j = (c_{j0}, \ldots, c_{jd})^T$, $\mathbf{d}_j = (d_{j0}, \ldots, d_{jd})^T$, and $K_{h_1}(.) = K(./h_1)/h_1$ is a re-scaled kernel function with a bandwidth h_1. Denote by $\{\hat{\mathbf{c}}_j, \hat{\mathbf{d}}_j\} = \text{argmin}_{\mathbf{c}_j, \mathbf{d}_j} \ell_j(\mathbf{c}_j, \mathbf{d}_j)$ the resulting local linear estimators. It is clearly known that $\hat{\alpha}_{jl} = \hat{c}_{jl}$ for $l = 0, \ldots, d$. Under certain regularity conditions, the resulting estimate $\hat{\alpha}_{jl}(U)$ is a consistent estimate of $\alpha_{jl}(U)$ with the nonparametric convergent rate (Fan & Zhang (1999)). Therefore, a consistent estimate of $\mathbb{E}(Y_j|\mathbf{x}, U)$ is

$$\hat{\mathbb{E}}(Y_j|\mathbf{x}, U) = \hat{\alpha}_{jl}(U) + \mathbf{x}^T \hat{\boldsymbol{\alpha}}_j(U), \text{ for } j = 1, \ldots, m.$$

4.2 Sampling Properties

In subsection 4.1.2 it is assumed that the conditional mean $\mathbb{E}(\mathbf{y}|\mathbf{x}, U)$ is known for the ease of illustration. In subsection 4.1.3 the estimation for $\mathbb{E}(\mathbf{y}|\mathbf{x}, U)$ is discussed when the sample $\{(U_i, \mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ is observable. Therefore, the residuals can be ob-

tained through replacing $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$ with its consistent estimation $\hat{\mathbb{E}}(\mathbf{y}|\mathbf{x},U)$, which can be noted as follows,

$$\hat{\varepsilon}_{ij} = Y_{ij} - \left\{ \hat{\alpha}_{j0}(U_i) + \hat{\boldsymbol{\alpha}}_j^T \mathbf{x}_i \right\}, \tag{4.26}$$

for $i = 1,\ldots,n$ and $j = 1,\ldots,m$. Then we use the residuals $\hat{\varepsilon}_{ij}$'s to play the role of the error ε_{ij}'s in subsection 4.1.2. Accordingly $\beta_{kj}(U)$'s, γ_{kjl}'s, and $\sigma_k^2(U)$'s can be estimated by using (4.18), (4.19), and (4.22). Subsequently, (4.3) and (4.21) can by adopted to estimate the conditional covariance matrix $\boldsymbol{\Sigma}(\mathbf{x},U) = \mathrm{Cov}(\boldsymbol{\varepsilon}|\mathbf{x},U)$. With slightly notational abuse, hereafter we still use the notations $\hat{\beta}_{kj}(U)$'s, $\hat{\gamma}_{kjl}$'s, and $\hat{\sigma}_k^2(U)$'s to denote the estimators of $\beta_{kj}(U)$'s, γ_{kjl}'s, and $\sigma_k^2(U)$'s, when the errors are replaced with the corresponding residuals. With such alternate, it is of interest to investigate how the estimation error due to the estimate of $\mathbb{E}(\mathbf{y}|\mathbf{x},U)$ affects the estimation of the conditional covariance matrix $\boldsymbol{\Sigma}(\mathbf{x},U)$. To address this issue, the following theorems derive the asymptotic bias and the variance of $\hat{\beta}_{kj}(U)$'s, $\hat{\gamma}_{kjl}$'s, and $\sigma_k^2(U)$'s, respectively, when the error ε_{ij}'s are replaced by the residuals $\hat{\varepsilon}_{ij}$'s.

The following technical conditions are imposed. We remark here that these are not the weakest possible conditions, but they are imposed to facilitate the proofs.

(C1). (The density of the index variable) Suppose U has a compact support and a probability density $f(\cdot)$, bounded away from 0 and with continuous derivative.

(C2). (The Kernel function) Suppose that the kernel function $K(\cdot)$ is a symmetric density function with a compact support.

(C3). (Smoothness of relevant functions) Assume that (i) the functions $\mathbb{E}(\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T|U)$ and $\mathbb{E}\{(\boldsymbol{\varepsilon}\otimes\mathbf{x})(\boldsymbol{\varepsilon}\otimes\mathbf{x})^T(\boldsymbol{\varepsilon}\otimes\mathbf{x})(\boldsymbol{\varepsilon}\otimes\mathbf{x})^T|U\}$ are continuous; (ii) Suppose that the functions $\boldsymbol{\alpha}(U)$, $\boldsymbol{\beta}(U)$, $\sigma_k^2(U)$, $\mathbb{E}(\mathbf{x}\mathbf{x}^T|U)$, and $\mathbb{E}\{(\boldsymbol{\varepsilon}\otimes\mathbf{x})(\boldsymbol{\varepsilon}\otimes\mathbf{x})^T|U\}$ have continuous third derivatives.

(C4). (The bandwidth) Suppose that the bandwidths satisfies $nh_n/(\log^2 n) \to \infty$ as $h_n \to 0$; $nh_1/(\log^2 n) \to \infty$ as $h_1 \to 0$; and $nh_2/(\log^2 n) \to \infty$ as $h_2 \to 0$.

(C5). (The moment requirement) Suppose the covariate vector $\mathbf{x}$ and the error vectors $\boldsymbol{\varepsilon}$ and $\mathbf{e}$ satisfy that $\mathbb{E}(\|\mathbf{x}\|^4|U) < \infty$, $\mathbb{E}(\|\mathbf{e}\|^{2+\delta}|U) < \infty$, and $\mathbb{E}(\|\boldsymbol{\varepsilon}\|^{2+\delta}|U) < \infty$ for some $\delta > 0$.

Let $\mu_i = \int u^i K(u)du$, $v_i = \int u^i K^2(u)du$, and $c_n(h) = h^2 + \log n/(nh)^{1/2}$ for an arbitrary bandwidth h.

Theorem 4.2.1. *Suppose conditions (C1) - (C5) hold true. Then*

$$(nh_1)^{1/2}\left\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U) - \boldsymbol{\beta}_{k,(k-1)}(U) - \tfrac{1}{2}h_1^2\mu_2\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U)\right\} \\ \to_D N\left(\mathbf{0}, v_0\{var(\boldsymbol{\varepsilon}|U)\}^{-1}\frac{\sigma_k^2(U)}{f(U)}\right), \tag{4.27}$$

for $k = 2,3,\ldots,m$, where $\mu_2 = \int u^2K(u)du$ and $v_0 = \int K^2(u)du$ and $f(U)$ denotes the density function of covariate U.

Theorem 4.2.2. *Suppose conditions (C1) - (C5) hold true. Then the estimator of $\boldsymbol{\gamma}_k$ is asymptotically normal, that is,*

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) \quad \to_D \quad N\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma\right), \tag{4.28}$$

for $k = 2,3,\ldots,m$, and $\boldsymbol{\Sigma}_\gamma$ is as follows:

$$\begin{aligned}\boldsymbol{\Sigma}_\gamma \;=\; & \sigma_k^2(\mathbf{x},U)\Big[\mathbb{E}\left\{(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})^T\right\} \\ & -\mathbb{E}\left\{\mathbb{E}\left\{(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})\mathbf{x}^T|U\right\}\mathbb{E}(\mathbf{x}\mathbf{x}^T|U)^{-1}\mathbb{E}\left\{\mathbf{x}(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})^T|U\right\}\right\}\Big]^{-1}.\end{aligned}$$

Theorem 4.2.3. *Suppose conditions (C1) - (C5) hold true. Then*

$$(nh_2)^{1/2}\left\{\hat{\sigma}_k^2(U) - \sigma_k^2(U) - bias\right\} \quad \to_D \quad N\left(0, v_0\frac{var(e_k^2|U)}{f(U)}\right), \tag{4.29}$$

for $k = 2,3,\ldots,m$, where $bias = h_2^2\mu_2\left\{\frac{\dot{\sigma}_k^2\dot{f}(U)}{f(U)} + \frac{\ddot{\sigma}_k^2(U)}{2}\right\}$, $\dot{f}(U)$, $\dot{\sigma}_k^2(U)$ and $\ddot{\sigma}_k^2(U)$ denotes the gradients and Hessian, respectively.

4.3 Numerical Studies

4.3.1 Simulation Studies

In the following simulation studies, it is assumed that $m = 6$ and $d = 3$. The simulated data sets are generated as follows. The covariate vector $\mathbf{x}$ follows from a multivariate

normal distribution with mean 0 and covariance matrix $(\rho_{ij})_{3\times 3}$, where $\rho_{ij} = \rho^{|i-j|}$. Moreover, the parameter ρ is set to be $0.25, 0.50$, and 0.75 corresponding to low, median, and high correlations among the covariate vector $\mathbf{x}$, which shows the effects of collinearity on the estimation of covariance matrix. In addition, the covariate U is taken from a uniform distribution on the interval $[0,1]$.

Next, let us generate the parameters and smoothing functions in the partially linear models. To be precise, the errors of the random error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)$, which are e_k's with $k = 1, \ldots, 6$, are generated independently from normal population with mean 0 and conditional variance function $\sigma_k^2(U)$, where $\sigma_k^2(U) = \sigma_0^2\{0.1 + |s_k(U)|\}^2$ given U values. To examine the effects of different noise levels, σ_0 is set to be $1, 2$, and 4. The elements of $\boldsymbol{\gamma}_{kj}$'s in model (4.6) are selected as follows:

$$
\begin{aligned}
\boldsymbol{\gamma}_{21} &= (0.1881, -0.0792, -0.0396)^T, \\
\boldsymbol{\gamma}_{31} &= (-0.1980, 0.0495, -0.1287)^T, \\
\boldsymbol{\gamma}_{32} &= (-0.4554, -0.0990, -0.4950)^T, \\
\boldsymbol{\gamma}_{41} &= (0.4158, -0.3069, 0.0495)^T, \\
\boldsymbol{\gamma}_{42} &= (0.3465, 0.4059, 0.4356)^T, \\
\boldsymbol{\gamma}_{43} &= (-0.1287, 0.0693, -0.1683)^T, \\
\boldsymbol{\gamma}_{51} &= (0.1287, 0.3366, 0.2178)^T, \\
\boldsymbol{\gamma}_{52} &= (0.1980, -0.1287, 0.0099)^T, \\
\boldsymbol{\gamma}_{53} &= (-0.0990, -0.0792, 0.2772)^T, \\
\boldsymbol{\gamma}_{54} &= (-0.0891, 0.0990, -0.0099)^T, \\
\boldsymbol{\gamma}_{61} &= (0.1980, -0.3861, 0.3267)^T, \\
\boldsymbol{\gamma}_{62} &= (0.4851, 0.1683, 0.1782)^T, \\
\boldsymbol{\gamma}_{63} &= (0.3069, -0.1386, 0.4950)^T, \\
\boldsymbol{\gamma}_{64} &= (0.2079, -0.3564, 0.4653)^T, \\
\boldsymbol{\gamma}_{65} &= (-0.0198, 0.0693, -0.4455)^T.
\end{aligned}
$$

The coefficients $\alpha_{jl}(U)$'s, $\beta_{kj}(U)$'s, and $s_k^2(U)$'s, are randomly selected from the

following functions:

$$\frac{1}{6}\left\{\frac{3}{2}U^{1/2}, 2U, -3U^2, -4U^3, \frac{\pi}{2}\cos(\frac{\pi U}{2}), \frac{\pi}{2}\sin(\frac{\pi U}{2}), \right.$$
$$\left.\frac{-\exp(U)}{\exp(1)-1}, \frac{\log(1+U)}{2\log(2)-1}, \frac{(1-U)/(1+U)}{2\log(2)-1}, \frac{6U(1+U)}{5}.\right\}$$

To examine the performance of our method, we repeat the experiment 1000 times with sample size $n = 400, 600$, and 800.

In order to investigate the performance of proposed methods, two scenarios are considered. The first scenario adopts the estimated errors $\hat{\varepsilon}_{ij}$'s to access the estimation, which is referred to as the "residual" estimator. Therefore, the residuals are estimated by

$$\begin{aligned}\hat{e}_{i1} &= \hat{\varepsilon}_{i1}, \\ \hat{e}_{ik} &= \hat{\varepsilon}_{ik} - \sum_{j=1}^{k-1}\{\hat{\beta}_{kj}(U) + \hat{\boldsymbol{\gamma}}_{kj}^T \mathbf{x}_i\}\hat{\varepsilon}_{ij},\end{aligned}$$

for $i = 1, \ldots, n$ and $k = 2, \ldots, m$. In the second scenario, we use the true errors ε_{ij}'s, which is referred to as the "error" estimator.

First, we examine the performance of our proposed method in estimating conditional covariance matrix $\Sigma(\mathbf{x}, U)$. Denote by $\hat{\Sigma}(\mathbf{x}, U)$ the estimate of $\Sigma(\mathbf{x}, U)$. The following two criteria are used to evaluate the estimation accuracy of $\hat{\Sigma}(\mathbf{x}, U)$:

$$\triangle_1 = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{tr}\{\mathbf{A}(\mathbf{x}_i, U_i)\} - \log|\mathbf{A}(\mathbf{x}_i, U_i)|] - m, \text{ and} \tag{4.30}$$

$$\triangle_2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}\{\mathbf{A}(\mathbf{x}_i, U_i) - \mathbf{I}\}^2, \tag{4.31}$$

$$\tag{4.32}$$

where $\mathbf{A}(\mathbf{x}_i, U_i) = \Sigma^{-1}(\mathbf{x}_i, U_i)\hat{\Sigma}(\mathbf{x}_i, U_i)$, $\mathrm{tr}(\mathbf{A}(\mathbf{x}_i, U_i))$ is the trace of matrix $\mathbf{A}(\mathbf{x}_i, U_i)$ and $|\mathbf{A}(\mathbf{x}_i, U_i)|$ is its determinant. The two losses $\triangle_1$ and $\triangle_2$ are usually referred to as Stein loss and the quadratic loss in the literature, for example, Muirhead (1982).

Table 4.1 and 4.2 summarize respectively the average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of

Stein loss $\triangle_1$ and the quadratic loss $\triangle_2$ over 1000 repetitions. It can be seen from Table 4.1 and 4.2 that the proposed method performs better with the increase of the sample size n. With different σ_0 levels, the proposed method performs similarly, which indicates that the proposed method is very robust to the noise level σ_0. In addition, it is noted that the correlation coefficient ρ has negative effects on the performance results. Specifically, the smaller the correlation is the better the proposed method performs. Comparing the results using "residuals" and that using "errors", it is not surprising to see that using "residuals" estimator performs relatively to that using "error" estimator.

Next, we assess the performance of our proposed method in estimating $\boldsymbol{\alpha}(U)$ via the mean squared errors (MSE$_{\boldsymbol{\alpha}}$), which is defined as follows:

$$\text{MSE}_{\boldsymbol{\alpha}} = \frac{1}{24n}\sum_{i=1}^{n}\sum_{j=1}^{6}\sum_{l=0}^{3}\left\{\hat{\alpha}_{jl}(U_i) - \alpha_{jl}(U_i)\right\}^2. \tag{4.33}$$

Table 4.3 summaries respectively the average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of MSE$_{\boldsymbol{\alpha}}$ over 1000 repetitions. It can be seen from Table 4.3 that the local linear approximation offers an accurate estimate for the mean functions. The proposed method performs better in estimating the conditional mean functions with the increase of the sample size n, and deteriorates with the increase of σ_0.

Similarly, the performance of the proposed method in estimating $\boldsymbol{\beta}(U)$,$\boldsymbol{\gamma}$, and $\sigma_k^2(U)$ are evaluated via the mean squared errors MSE$_{\boldsymbol{\beta}}$, MSE$_{\boldsymbol{\gamma}}$, and MSE$_{\boldsymbol{\sigma}}$ respectively. Specifically,

$$\text{MSE}_{\boldsymbol{\beta}} = \frac{1}{5n}\sum_{i=1}^{n}\sum_{k=2}^{6}\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\left\{\hat{\beta}_{kj}(U_i) - \beta_{kj}(U_i)\right\}^2\right], \tag{4.34}$$

$$\text{MSE}_{\boldsymbol{\gamma}} = \frac{1}{5}\sum_{k=2}^{6}\sum_{l=1}^{3}\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\left\{\hat{\gamma}_{kj} - \gamma_{kj}\right\}^2\right], \tag{4.35}$$

$$\text{MSE}_{\sigma} = \frac{1}{6n}\sum_{i=1}^{n}\sum_{k=1}^{6}\left\{\hat{\sigma}_k^2(U_i) - \sigma_k^2(U_i)\right\}^2. \tag{4.36}$$

Table 4.4, Table 4.5, and Table 4.6 summarize average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of

$\text{MSE}_{\boldsymbol{\beta}}$, $\text{MSE}_{\boldsymbol{\gamma}}$, and MSE_{σ} over 1000 repetitions, respectively. It can be observed from Table 4.4 that in estimating $\boldsymbol{\beta}(U)$ our propose using "residuals" behaves very similarly to that using "errors", though the latter performs slightly better in most cases. This confirms the theoretical investigation in Theorem. In addition, the proposed method is very robust to the noise level σ_0, which can be found analytically in view of (4.7). Similarly to the comments on estimates $\boldsymbol{\beta}(U)$, it can be observed similar phenomenon in estimating the $\boldsymbol{\gamma}$ in Table 4.5. In Table 4.6, the estimate of $\sigma_k^2(U)$ deteriorates with the increase of σ_0, which parallels to the observation in estimating $\boldsymbol{\alpha}(U)$. The proposed method using "errors" performs slightly better than that using "residuals".

Figure 4.1 and 4.2 depict the boxplots for the Stein loss $\triangle_1$ and quadratic loss $\triangle_2$ over 1000 iterations respectively with the sample size n selected as 400,600, and 800, the correlation coefficient ρ selected as 0.25,0.50, and 0.75, the noise σ_0 specified as 1. Here the boxplot is used to graphically depict data through five-number summaries: the smallest observation, lower quartile, median, upper quartile, and largest observation. Corresponding to the discussion on the estimation accuracy of $\hat{\boldsymbol{\Sigma}}(\mathbf{x},U)$ from Table 4.1 and 4.2, it can be observed that with the increase of the sample size n the median of two loss functions decrease, which thus indicate the proposed method performs better. In addition, the performance deteriorates with the increase of correlation coefficient ρ. The boxplots also indicate some observations may be outliers. Figure 4.3 to Figure 4.6 depict the histograms for$\text{MSE}_{\boldsymbol{\alpha}}$, $\text{MSE}_{\boldsymbol{\beta}}$, $\text{MSE}_{\boldsymbol{\gamma}}$, and MSE_{σ} with the noise level σ_0 specified as 1.

Figure 4.7 depict the estimated conditional mean function with varying coefficient models and their 95% pointwise confidence intervals for a simulation study with the sample size $n = 800$, the correlation coefficient $\rho = 0.25$, and the noise level $\sigma_0 = 1$. The black solid lines are the true mean functions $\boldsymbol{\alpha}_{ij}(U)$, for $i = 1,\ldots,6,\ j = 0,\ldots,3$; the blue dash-dotted lines are the estimated mean functions $\hat{\boldsymbol{\alpha}}_{ij}(U)$, for $i = 1,\ldots,6,\ j = 0,\ldots,3$; and the red dashed lines are the corresponding 95% pointwise confidence intervals. Figure 4.7 demonstrates that the estimated conditional mean functions are accurate approximations to the true conditional mean functions. Accordingly, Figure 4.8 depict the estimated $\boldsymbol{\beta}_{kj}(U)$ with $k = 2,\ldots,6,\ j = 1,\ldots,k-1$ in model (4.6) and their 95% pointwise confidence intervals for a simulation study with the sample size $n = 800$, the correlation coefficient $\rho = 0.25$, and the noise level $\sigma_0 = 1$. It is not surprising to see that the proposed method produce good approximations to the true smoothing functions.

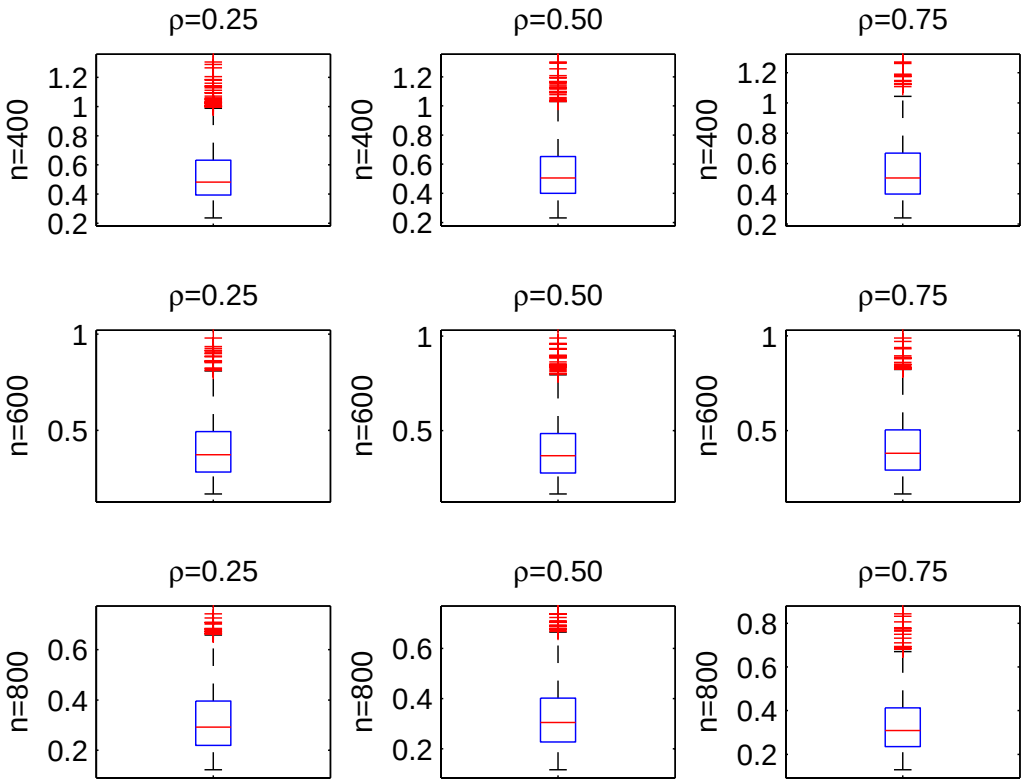

Figure 4.1. Boxplot for Stein loss $\triangle_1$ given $\sigma_0 = 1$.

4.3.2 Real Data Application

We get most of our energy from nonrenewable energy sources, of which the three most often used energy sources are: crude oil (also known as petroleum), natural gas, and coal. Coal is the most abundant fossil fuel produced in the United States and is relatively inexpensive to produce and convert to useful energy. It is known as a nonrenewable energy source because it takes millions of years to create. However, producing and using coal has many impacts on the environment. In the United States, most coal is used as a fuel to generate electricity. Burning coal produces numerous emissions that adversely affect the environment and human health. Crude oil is refined into the petroleum products that are burned to produce energy, they may be used to propel a vehicle, to heat a building, or to produce electric power in a generator. Natural gas has many qualities that make it an efficient, relatively clean, and economical energy source. Burning natural

Table 4.1. The performance of proposed functional estimation of conditional covariance matrix $\Sigma(\mathbf{x}, U)$ via Stein loss $\triangle_1$. The average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of Stein loss $\triangle_1$ over 1000 repetitions. The numbers in the last six columns are multiplied by a factor of 100.

			residual			error		
			σ_0			σ_0		
sample size	ρ		1.0	2.0	4.0	1.0	2.0	4.0
400	25	aver.	53.15	52.89	53.08	36.51	36.85	36.68
		stdev.	18.55	18.65	18.72	8.51	8.90	8.76
		median	48.23	47.78	48.78	34.94	35.36	35.25
		mad.	11.00	10.79	11.26	5.19	5.46	5.43
	50	aver.	54.58	54.19	52.76	36.97	36.92	36.72
		stdev.	19.25	18.26	18.43	8.66	8.53	9.11
		median	50.54	50.69	48.96	35.55	35.66	35.40
		mad.	11.69	11.26	11.70	5.73	5.57	5.15
	75	aver.	54.80	55.95	55.32	36.37	37.07	36.70
		stdev.	19.23	20.26	18.65	8.42	8.81	8.26
		median	50.53	51.05	51.05	35.25	35.45	35.40
		mad.	12.39	12.25	11.39	5.52	5.26	5.14
600	25	aver.	40.44	38.71	38.86	25.61	25.43	25.32
		stdev.	15.73	13.98	14.08	6.46	6.24	6.29
		median	37.32	35.92	35.99	24.55	24.20	24.32
		mad.	10.46	8.99	9.58	4.42	3.99	3.95
	50	aver.	39.91	39.08	39.07	25.41	25.24	25.02
		stdev.	15.67	15.05	14.36	6.57	6.58	6.13
		median	36.65	35.73	35.82	24.21	24.02	24.11
		mad.	10.03	9.49	9.03	4.26	4.03	3.82
	75	aver.	41.30	39.20	39.82	25.32	24.73	25.27
		stdev.	15.38	14.92	14.70	6.46	6.04	6.18
		median	38.07	35.48	37.08	24.16	23.58	24.22
		mad.	9.67	9.36	9.62	3.98	3.88	4.17
800	25	aver.	32.00	31.60	32.28	19.59	19.57	19.97
		stdev.	12.46	12.31	12.12	5.13	5.16	5.26
		median	29.22	29.29	30.30	18.55	18.64	19.17
		mad.	8.22	8.50	8.53	3.25	3.30	3.66
	50	aver.	32.81	32.70	32.02	19.83	19.99	19.80
		stdev.	12.67	12.80	11.88	5.10	5.16	4.97
		median	30.41	30.02	29.78	18.85	18.93	18.91
		mad.	8.55	8.85	8.13	3.31	3.28	3.28
	75	aver.	33.58	33.27	33.05	19.80	19.71	19.93
		stdev.	12.91	13.33	12.62	5.02	5.22	5.15
		median	30.90	30.73	30.65	18.98	18.64	18.94
		mad.	8.26	8.72	8.51	3.21	3.29	3.22

Table 4.2. The performance of proposed functional estimation of conditional covariance matrix $\Sigma(\mathbf{x}, U)$ via quadratic loss $\triangle_2$. The average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of quadratic loss $\triangle_2$ over 1000 repetitions.

			residual			error		
			σ_0			σ_0		
sample size	ρ		1.0	2.0	4.0	1.0	2.0	4.0
400	25	aver.	1.78	1.75	1.76	0.95	0.96	0.96
		stdev.	1.09	1.06	1.06	0.36	0.39	0.37
		median	1.40	1.41	1.45	0.85	0.86	0.86
		mad.	0.49	0.50	0.51	0.19	0.20	0.20
	50	aver.	1.84	1.82	1.75	0.97	0.97	0.96
		stdev.	1.08	1.04	1.03	0.37	0.36	0.41
		median	1.52	1.52	1.45	0.86	0.89	0.88
		mad.	0.55	0.53	0.53	0.21	0.21	0.20
	75	aver.	1.86	1.90	1.87	0.95	0.97	0.96
		stdev.	1.11	1.19	1.08	0.37	0.38	0.36
		median	1.53	1.56	1.57	0.86	0.86	0.88
		mad.	0.55	0.56	0.55	0.20	0.19	0.19
600	25	aver.	1.31	1.21	1.22	0.65	0.64	0.64
		stdev.	0.77	0.65	0.67	0.24	0.23	0.24
		median	1.11	1.05	1.04	0.60	0.58	0.59
		mad.	0.44	0.38	0.39	0.15	0.14	0.14
	50	aver.	1.27	1.24	1.23	0.64	0.64	0.63
		stdev.	0.75	0.72	0.69	0.24	0.25	0.23
		median	1.07	1.03	1.04	0.58	0.58	0.58
		mad.	0.42	0.40	0.38	0.15	0.13	0.13
	75	aver.	1.33	1.24	1.26	0.64	0.62	0.64
		stdev.	0.74	0.73	0.70	0.24	0.22	0.23
		median	1.14	1.02	1.09	0.58	0.56	0.59
		mad.	0.43	0.39	0.41	0.14	0.13	0.14
800	25	aver.	0.98	0.97	0.99	0.49	0.48	0.50
		stdev.	0.54	0.54	0.51	0.18	0.18	0.18
		median	0.83	0.84	0.88	0.44	0.44	0.46
		mad.	0.33	0.32	0.35	0.11	0.11	0.12
	50	aver.	1.02	1.01	0.98	0.49	0.50	0.49
		stdev.	0.56	0.56	0.51	0.18	0.18	0.17
		median	0.88	0.87	0.86	0.45	0.45	0.45
		mad.	0.34	0.35	0.32	0.11	0.11	0.11
	75	aver.	1.04	1.04	1.03	0.49	0.49	0.50
		stdev.	0.56	0.61	0.56	0.17	0.18	0.18
		median	0.90	0.88	0.88	0.45	0.44	0.45
		mad.	0.34	0.35	0.34	0.11	0.11	0.11

Table 4.3. The performance of proposed method in estimating $\boldsymbol{\alpha}(U)$ via mean squared errors $\text{MSE}_{\boldsymbol{\alpha}}$. The average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of $\text{MSE}_{\boldsymbol{\alpha}}$ over 1000 repetitions. The numbers in the last six columns are multiplied by a factor of 100.

			residual		
			σ_0		
sample size	ρ		1.0	2.0	4.0
400	25	aver.	0.53	2.07	8.30
		stdev.	0.19	0.76	2.94
		median	0.48	1.91	7.76
		mad.	0.10	0.42	1.60
	50	aver.	0.63	2.51	9.91
		stdev.	0.23	0.94	3.78
		median	0.59	2.32	9.04
		mad.	0.12	0.49	1.97
	75	aver.	1.08	4.20	17.38
		stdev.	0.43	1.69	6.88
		median	0.99	3.81	16.13
		mad.	0.24	0.88	3.93
600	25	aver.	0.35	1.38	5.43
		stdev.	0.12	0.49	1.98
		median	0.33	1.27	5.04
		mad.	0.07	0.27	1.06
	50	aver.	0.43	1.67	6.60
		stdev.	0.15	0.58	2.17
		median	0.40	1.56	6.14
		mad.	0.08	0.34	1.24
	75	aver.	0.70	2.79	11.26
		stdev.	0.26	1.06	4.57
		median	0.65	2.59	10.33
		mad.	0.15	0.59	2.54
800	25	aver.	0.27	1.03	4.19
		stdev.	0.09	0.34	1.49
		median	0.25	0.98	3.90
		mad.	0.05	0.19	0.83
	50	aver.	0.32	1.26	5.02
		stdev.	0.11	0.49	1.68
		median	0.30	1.13	4.66
		mad.	0.06	0.24	0.94
	75	aver.	0.54	2.15	8.34
		stdev.	0.20	0.84	2.91
		median	0.50	1.98	7.79
		mad.	0.11	0.46	1.72

Table 4.4. The performance of proposed method in estimating $\boldsymbol{\beta}$ via mean squared errors $\text{MSE}_{\boldsymbol{\beta}}$. The average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of $\text{MSE}_{\boldsymbol{\beta}}$ over 1000 repetitions. The numbers in the last six columns are multiplied by a factor of 100.

			residual			error		
			σ_0			σ_0		
sample size	ρ		1.0	2.0	4.0	1.0	2.0	4.0
400	25	aver.	1.50	1.44	1.50	1.63	1.59	1.65
		stdev.	0.73	0.66	0.69	0.88	0.84	0.82
		median	1.31	1.29	1.34	1.38	1.35	1.42
		mad.	0.33	0.33	0.36	0.39	0.36	0.42
	50	aver.	1.45	1.42	1.49	1.59	1.52	1.63
		stdev.	0.66	0.60	0.75	0.81	0.67	0.88
		median	1.28	1.26	1.30	1.38	1.34	1.37
		mad.	0.34	0.32	0.38	0.38	0.36	0.40
	75	aver.	1.45	1.53	1.56	1.52	1.60	1.65
		stdev.	0.66	0.69	0.73	0.79	0.81	0.83
		median	1.29	1.36	1.39	1.30	1.38	1.41
		mad.	0.32	0.36	0.39	0.35	0.38	0.42
600	25	aver.	0.92	0.91	0.94	1.03	1.01	1.05
		stdev.	0.42	0.40	0.43	0.50	0.48	0.54
		median	0.83	0.81	0.84	0.90	0.88	0.91
		mad.	0.23	0.20	0.21	0.25	0.23	0.25
	50	aver.	0.94	0.94	0.94	1.03	1.05	1.04
		stdev.	0.42	0.44	0.40	0.52	0.54	0.49
		median	0.83	0.83	0.84	0.89	0.91	0.91
		mad.	0.21	0.21	0.22	0.24	0.25	0.25
	75	aver.	0.97	0.93	0.95	1.03	1.01	1.03
		stdev.	0.46	0.42	0.45	0.53	0.52	0.52
		median	0.86	0.84	0.84	0.89	0.87	0.89
		mad.	0.22	0.21	0.22	0.25	0.24	0.23
800	25	aver.	0.69	0.67	0.69	0.77	0.76	0.78
		stdev.	0.34	0.32	0.32	0.42	0.40	0.40
		median	0.60	0.59	0.60	0.66	0.65	0.68
		mad.	0.16	0.15	0.17	0.18	0.18	0.19
	50	aver.	0.68	0.69	0.67	0.75	0.78	0.76
		stdev.	0.32	0.33	0.30	0.37	0.40	0.38
		median	0.61	0.60	0.60	0.65	0.66	0.65
		mad.	0.16	0.16	0.16	0.18	0.18	0.17
	75	aver.	0.72	0.71	0.70	0.78	0.78	0.78
		stdev.	0.35	0.32	0.32	0.41	0.38	0.39
		median	0.63	0.63	0.62	0.67	0.67	0.67
		mad.	0.18	0.16	0.16	0.20	0.18	0.19

Table 4.5. The performance of proposed method in estimating $\boldsymbol{\gamma}$ via mean squared errors $\text{MSE}_{\boldsymbol{\gamma}}$. The average ("average"), the standard deviation ("stdev."), the median ("median") and the median absolute deviation ("mad.") of $\text{MSE}_{\boldsymbol{\gamma}}$ over 1000 repetitions. The numbers in the last six columns are multiplied by a factor of 100.

			residual			error		
			σ_0			σ_0		
sample size	ρ		1.0	2.0	4.0	1.0	2.0	4.0
400	25	aver.	0.37	0.38	0.37	0.32	0.33	0.32
		stdev.	0.11	0.11	0.11	0.09	0.10	0.10
		median	0.36	0.36	0.35	0.31	0.32	0.31
		mad.	0.07	0.07	0.07	0.06	0.07	0.06
	50	aver.	0.49	0.48	0.48	0.43	0.42	0.43
		stdev.	0.15	0.15	0.16	0.14	0.14	0.14
		median	0.47	0.46	0.45	0.41	0.39	0.40
		mad.	0.09	0.10	0.10	0.08	0.08	0.09
	75	aver.	0.86	0.89	0.85	0.76	0.78	0.76
		stdev.	0.28	0.31	0.29	0.26	0.28	0.27
		median	0.82	0.85	0.80	0.72	0.74	0.72
		mad.	0.17	0.19	0.17	0.16	0.17	0.15
600	25	aver.	0.24	0.23	0.23	0.21	0.21	0.21
		stdev.	0.07	0.07	0.07	0.07	0.06	0.06
		median	0.22	0.22	0.22	0.20	0.20	0.20
		mad.	0.05	0.05	0.04	0.04	0.04	0.04
	50	aver.	0.31	0.30	0.30	0.27	0.27	0.27
		stdev.	0.10	0.10	0.09	0.09	0.09	0.08
		median	0.29	0.29	0.28	0.27	0.26	0.25
		mad.	0.06	0.06	0.05	0.05	0.05	0.05
	75	aver.	0.55	0.53	0.54	0.49	0.48	0.49
		stdev.	0.18	0.18	0.18	0.17	0.17	0.17
		median	0.52	0.50	0.51	0.46	0.46	0.46
		mad.	0.11	0.11	0.11	0.10	0.11	0.10
800	25	aver.	0.17	0.16	0.17	0.15	0.15	0.15
		stdev.	0.05	0.05	0.05	0.04	0.04	0.05
		median	0.16	0.16	0.16	0.14	0.15	0.14
		mad.	0.03	0.03	0.03	0.03	0.03	0.03
	50	aver.	0.22	0.22	0.21	0.20	0.20	0.20
		stdev.	0.07	0.07	0.07	0.06	0.06	0.06
		median	0.21	0.21	0.21	0.19	0.19	0.19
		mad.	0.04	0.04	0.04	0.04	0.04	0.04
	75	aver.	0.39	0.39	0.40	0.36	0.36	0.37
		stdev.	0.13	0.13	0.14	0.12	0.12	0.14
		median	0.37	0.37	0.37	0.34	0.34	0.34
		mad.	0.08	0.08	0.09	0.08	0.07	0.08

Table 4.6. The performance of proposed method in estimating σ via mean squared errors MSE_σ. The average ("average"), the standard deviation ("stdev."), the median ("median') and the median absolute deviation ("mad.") of MSE_σ over 1000 repetitions. The numbers in the last six columns are multiplied by a factor of 100.

			residual			error		
			σ_0			σ_0		
sample size	ρ		1.0	2.0	4.0	1.0	2.0	4.0
400	25	aver.	0.13	1.99	33.01	0.10	1.49	24.09
		stdev.	0.12	1.85	30.05	0.09	1.40	21.76
		median	0.09	1.36	22.76	0.07	1.02	17.07
		mad.	0.06	0.80	13.95	0.04	0.62	9.74
	50	aver.	0.13	1.99	33.61	0.10	1.50	24.84
		stdev.	0.11	1.75	28.59	0.08	1.39	21.55
		median	0.09	1.49	24.53	0.07	1.10	17.71
		mad.	0.06	0.89	14.69	0.04	0.63	10.35
	75	aver.	0.13	2.09	32.55	0.10	1.49	24.45
		stdev.	0.11	1.81	26.51	0.08	1.34	20.64
		median	0.09	1.58	24.18	0.07	1.05	17.55
		mad.	0.06	0.97	14.93	0.04	0.62	9.96
600	25	aver.	0.10	1.58	25.74	0.07	1.04	17.50
		stdev.	0.09	1.36	21.04	0.06	0.89	14.06
		median	0.07	1.15	19.19	0.05	0.75	13.08
		mad.	0.04	0.66	11.09	0.03	0.42	7.14
	50	aver.	0.11	1.61	25.26	0.07	1.08	17.50
		stdev.	0.09	1.31	22.99	0.06	0.89	16.19
		median	0.08	1.20	17.90	0.05	0.82	12.54
		mad.	0.05	0.74	10.77	0.03	0.45	6.93
	75	aver.	0.10	1.61	27.47	0.07	1.09	18.66
		stdev.	0.09	1.37	24.42	0.06	0.91	16.21
		median	0.07	1.19	19.73	0.05	0.82	14.04
		mad.	0.04	0.73	11.57	0.03	0.47	8.16
800	25	aver.	0.09	1.38	23.07	0.05	0.87	14.37
		stdev.	0.07	1.14	19.97	0.04	0.71	12.08
		median	0.06	0.97	17.36	0.04	0.64	10.69
		mad.	0.04	0.58	10.28	0.02	0.35	5.96
	50	aver.	0.08	1.33	22.23	0.05	0.88	13.89
		stdev.	0.07	1.15	18.98	0.05	0.75	11.79
		median	0.06	0.94	16.48	0.04	0.64	10.37
		mad.	0.03	0.52	9.94	0.02	0.37	5.99
	75	aver.	0.09	1.40	23.26	0.05	0.89	14.97
		stdev.	0.08	1.19	21.00	0.04	0.74	13.32
		median	0.06	1.03	15.89	0.04	0.67	10.70
		mad.	0.04	0.62	9.58	0.02	0.37	6.11

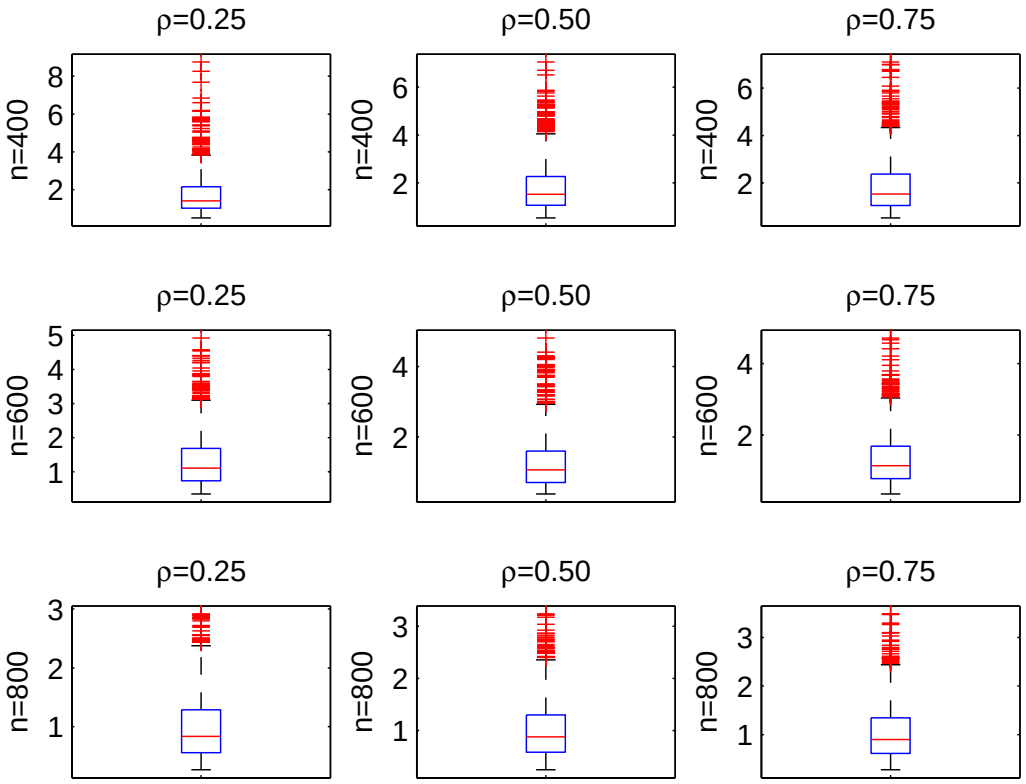

Figure 4.2. Boxplot for quadratic loss $\triangle_2$ given $\sigma_0 = 1$.

gas for energy results in much fewer emissions of nearly all types of air pollutants and carbon dioxide (CO_2) per unit of heat produced than coal or refined petroleum products.

In early years, the choices for most electric utility generators were large coal-powered plants. However, resulting from economic processes, technological innovations, and environmental developments , natural gas and crude oil have become the fuel of choice for new power plants due to their clean burning nature. It is of interest to investigate how the average electricity prices rely on the average fossil fuel costs.

We now illustrate the proposed method by an application to the Pennsylvania electricity load data set. The data set was collected at three distinct areas at Pennsylvania during the period of January through April in 2009. The power stations in these areas burn coal to generate electricity. When the demand of electricity load is beyond the supply of the current energy source, the power plants have to use other energy sources,

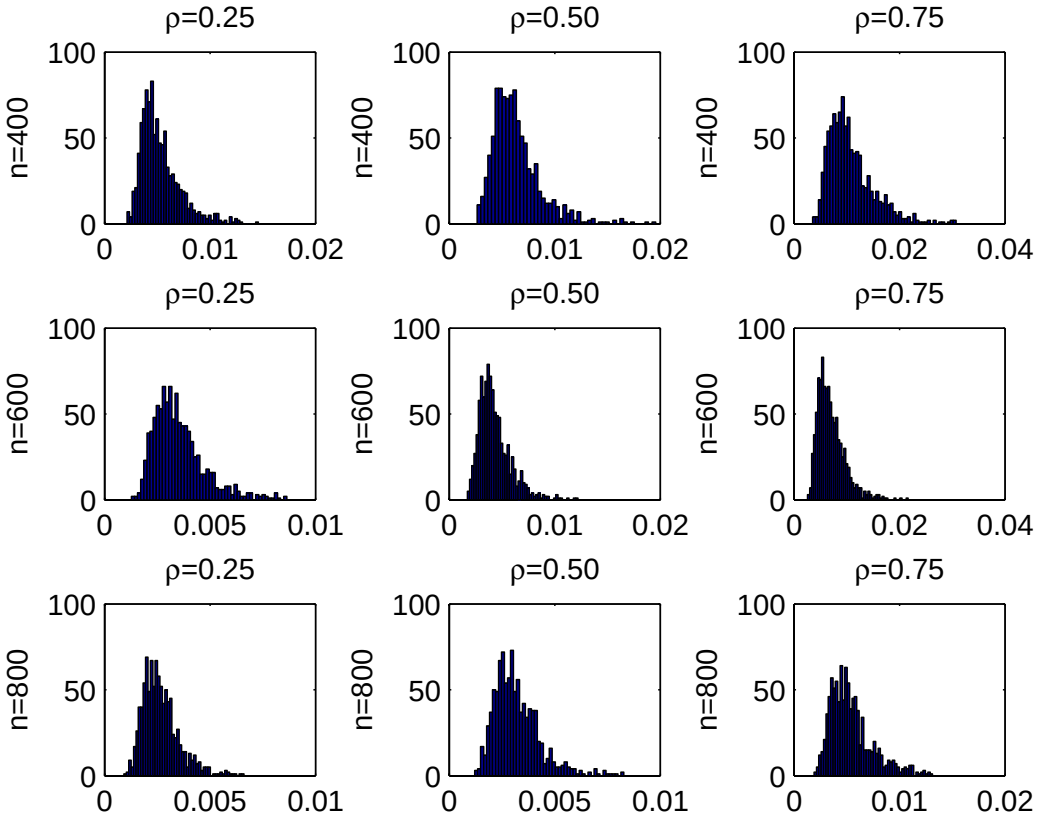

Figure 4.3. Histogram for mean square error $\text{MSE}_{\boldsymbol{\alpha}}$ given $\sigma_0 = 1$.

for example, crude oil and natural gas, to generate the electricity. Therefore, the average electricity prices may depend upon the average fossil fuel costs in very different ways. The data set consists of the average electricity prices at three areas of Pennsylvania, the overall electricity load, as well as the hourly prices of coal, crude oil and natural gas that might explain the variation in the average electricity prices. There are 2615 data points in total after removing a few observations with missing values. For simplicity of notation, the average electricity prices at these three areas of Pennsylvania is denoted by $\mathbf{y} = (Y_1, Y_2, Y_3)^T$, the covariate overall electricity load is denoted by U, and the covariates hourly prices of coal, crude oil and natural gas are denoted by $\mathbf{x} = (X_1, X_2, X_3)^T$. Note that in this application, the observations for the electricity load U is normalized so that its values are within interval $[0, 1]$. The objective of the study is to understand the association between the average electricity prices at these three areas of Pennsylvania

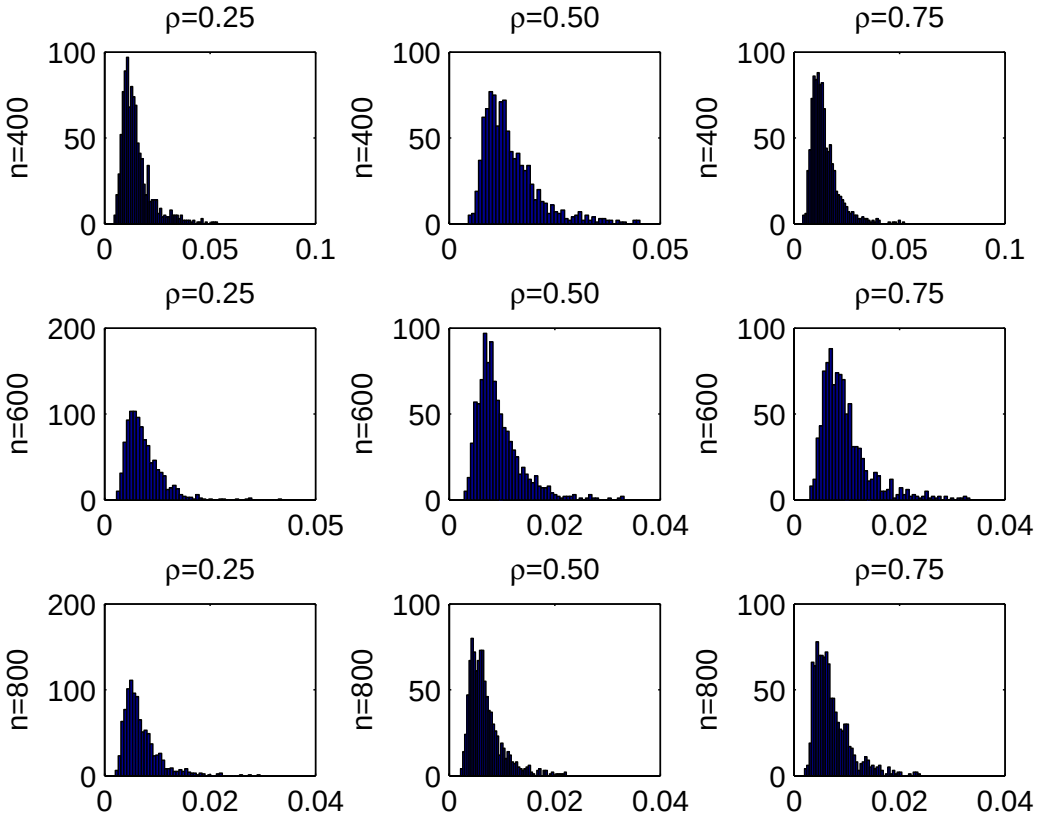

Figure 4.4. Histogram for mean square error $\mathrm{MSE}_{\boldsymbol{\beta}}$ given $\sigma_0 = 1$.

and the four covariates.

First, we investigate how the electricity rates change with the fossil fuel costs and the electricity load. The varying-coefficient model

$$Y_j = \alpha_{j0}(U) + \alpha_{j1}(U)X_1 + \alpha_{j2}(U)X_2 + \alpha_{j3}(U)X_3 + \varepsilon_j, \text{ for } j = 1,2,3, \qquad (4.37)$$

is fitted to the given data. Figure 4.9 presents the estimated varying-coefficient functions and their 95% pointwise confidence intervals. It demonstrates that when the electricity load is light, say U is less than 0.1, the average electricity prices rely on the hourly prices of crude oil in area one and area three. While the hourly prices of coal and natural gas have little effect on the average electricity prices at these three areas, since the relative varying-coefficient functions are not significantly different from zero within

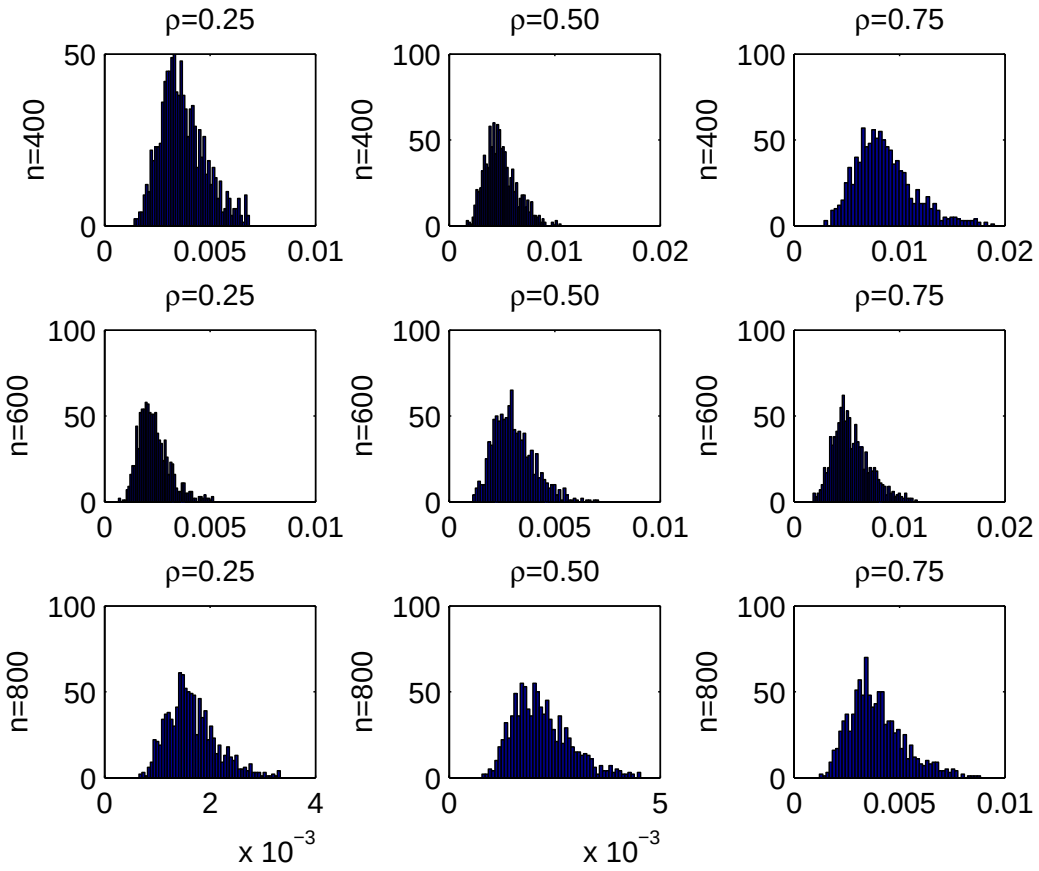

Figure 4.5. Histogram for mean square error $\text{MSE}_{\boldsymbol{\gamma}}$ given $\sigma_0 = 1$.

this interval. However, the fossil fuel prices affect the average prices in complicated patterns as the electricity load becomes heavier. To be specific, in area one, when the electricity load U is greater than 0.6, the average electricity prices are determined by all these three energy sources. As the U is close to 1, the hourly price of the coal and natural gas dominates the average electricity price seeing that the varying-coefficient functions are significantly varying from zero. In area two, the average electricity price depends upon the price of coal and natural gas when when U is larger than 0.8 and the price of coal dominates the average electricity price notably. While in area three, all these three energy sources influence the average price of electricity when the electricity load U is larger than 0.8 since all the corresponding varying-coefficient functions are significantly different from zero. Additionally, as the load U grows, the price of coal dominates the average electricity price considerably.

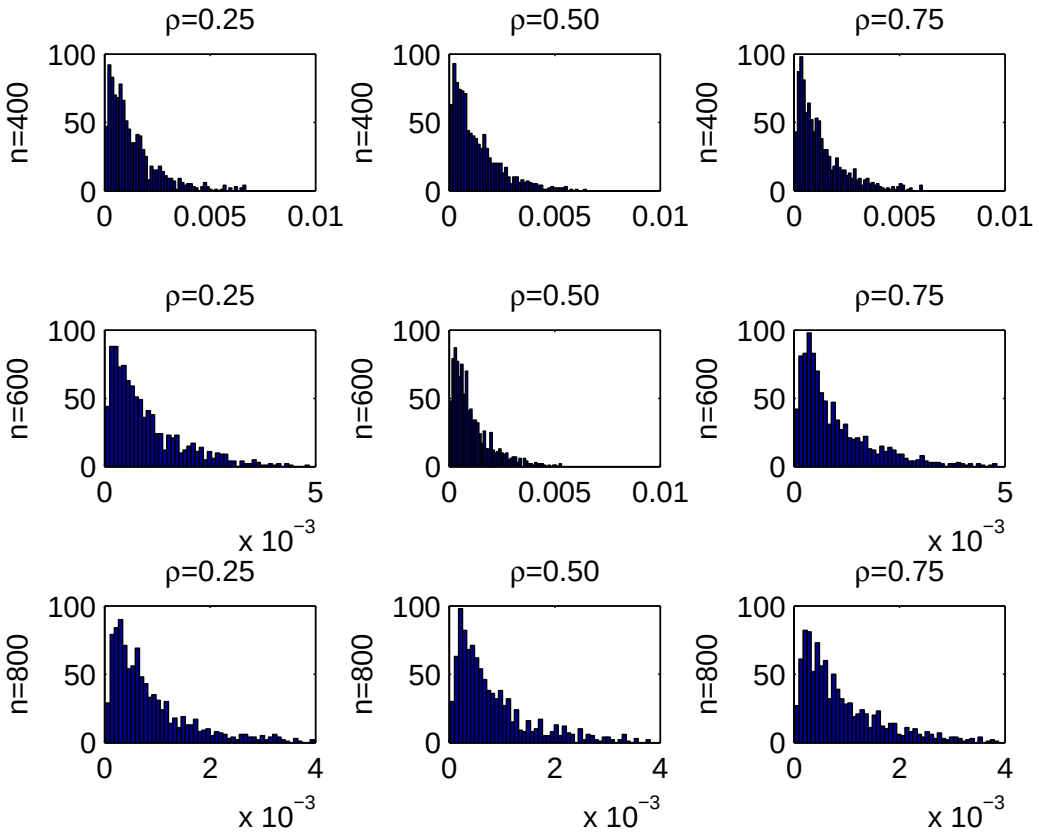

Figure 4.6. Histogram for mean square error MSE_σ given $\sigma_0 = 1$.

The above analysis presents the marginal effects the fossil fuel costs and the electricity load have on the average electricity prices in three areas of Pennsylvania. It is usually desirable to understand in advance the joint relationship amongst the average electricity prices in these three different districts. Therefore the study of the correlation structure of these three places is preferred. In order to achieve this aim, the proposed method via modified Cholesky decomposition is used to estimate the conditional covariance matrix. The residuals can be fitted through (4.5) and (4.6). Figure 4.10 depicts the estimated smoothing functions $\beta_{kj}(U)$'s in (4.6) and the corresponding 95% pointwise confidence intervals, where $k = 2,3$ and $j = 1,\ldots,k-1$. The blue dash-dotted lines are the estimated $\beta_{kj}(U)$'s, for $k = 2,3$ and $j = 1,\ldots,k-1$; the red dashed lines are the corresponding 95% pointwise confidence intervals. In addition, the estimated $\boldsymbol{\gamma}_{kj}$'s in

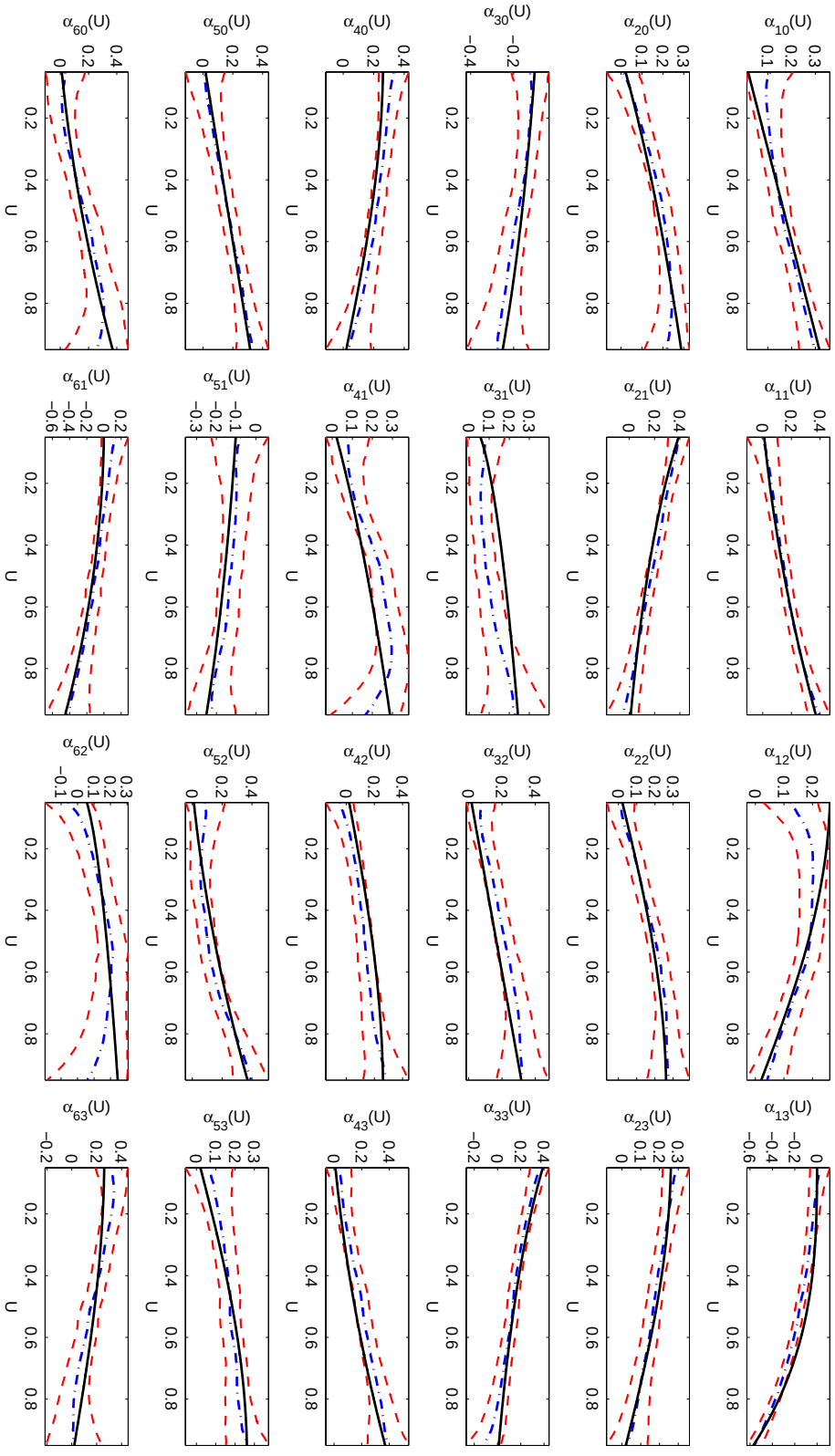

Figure 4.7. Estimated $\alpha_{ij}(U)$'s and their 95% pointwise confidence intervals, for $i = 1,\ldots,6,\ j = 0,\ldots,3$. The black solid lines are the true mean functions $\boldsymbol{\alpha}_{ij}(U)$; the blue dash-dotted lines are the estimated mean functions $\hat{\alpha}_{ij}(U)$; and the red dashed lines are the corresponding 95% pointwise confidence intervals. These legends remain the same for Figure 4.8.

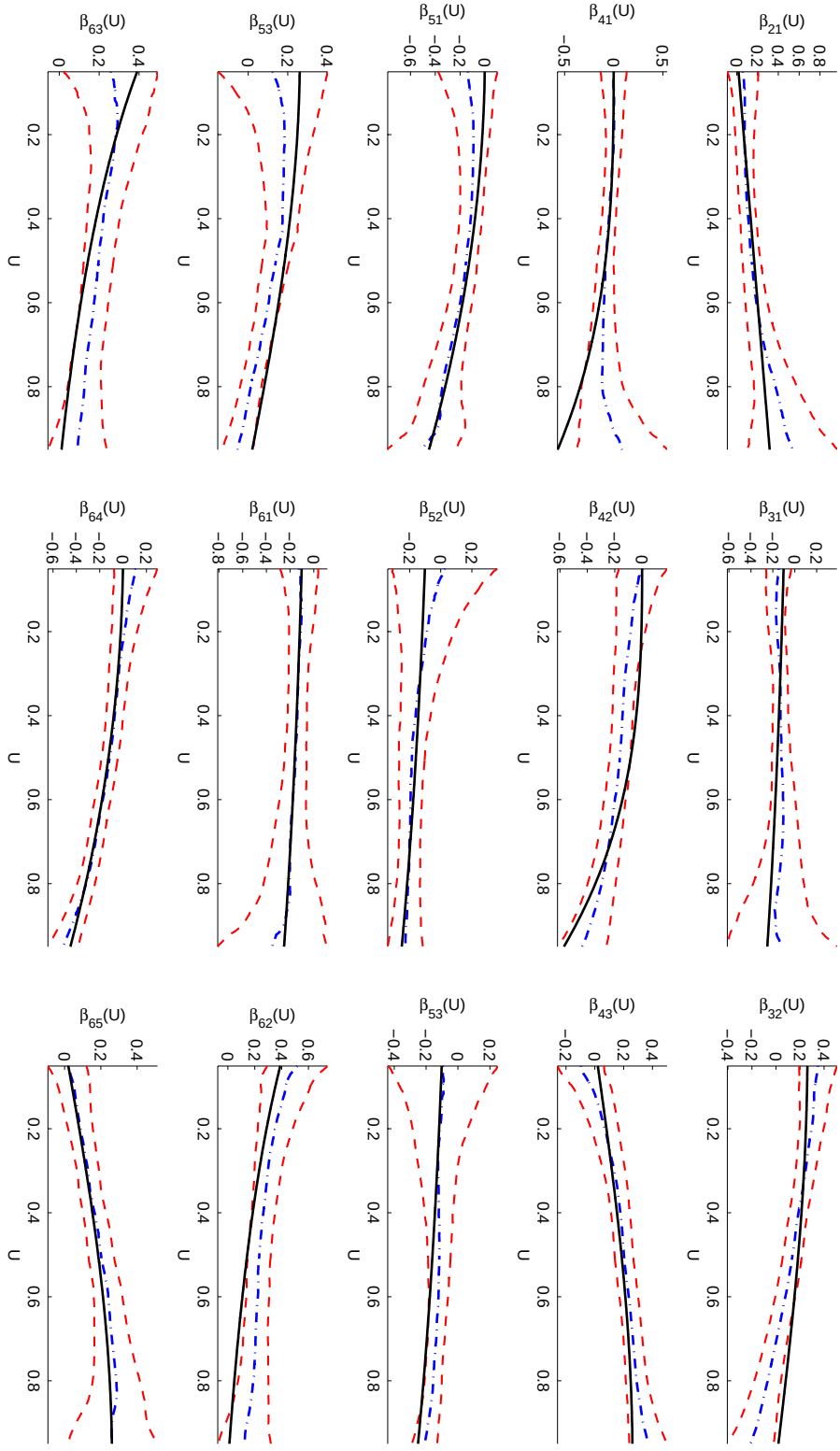

Figure 4.8. Estimated $\beta_{kj}(U)$'s and their 95% pointwise confidence intervals, for $k = 2,\ldots,6,\ j = 1,\ldots,k-1$.

(4.6) are as follows:

$$\begin{aligned}
\boldsymbol{\gamma}_{21} &= (0.2041, -0.4087, 0.1461)^T, \\
\boldsymbol{\gamma}_{31} &= (0.1592, -0.0685, 0.1198)^T, \\
\boldsymbol{\gamma}_{32} &= (-0.1289, 0.0486, -0.0832)^T.
\end{aligned}$$

The conditional correlation functions can be obtained through dividing the conditional covariance functions by the conditional variance functions. Figure 4.11 depicts the conditional correlation functions and their 95% pointwise confidence intervals when the covariate vector $\mathbf{x}$ is fixed at its center. It can be easily seen that the average electricity prices are positive-correlated at a high level when the electricity load U is relatively low. As the electricity load increases, the correlation structure changes with different trend. To be specific, when the U is close to 0.8, the correlation between the average electricity prices in area 1 and area 2 achieves its minimum. While there is a monotone decreasing trend for the corresponding correlation coefficient between area 1 and area 3. But even with high electricity load, these prices are still highly correlated since the correlation coefficient remain large, which is above 0.8. The correlation between the average electricity prices in area 2 and area 3 goes down as the electricity load increase. The different observation comparing to other two plots is that the correlation coefficient drop very quickly from 0.6 to 0.15, which indicating that with high electricity load, the average electricity prices in area 2 and area 3 is positive-correlated at a low level.

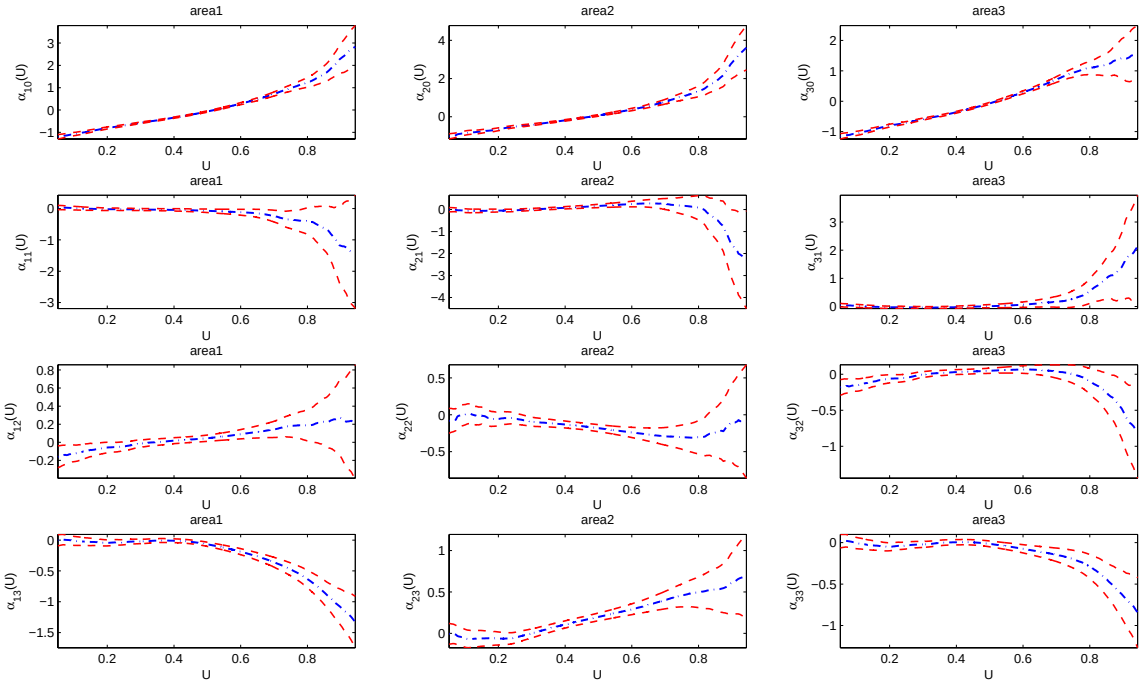

Figure 4.9. Estimated varying-coefficient $\alpha_{ij}(U)$'s and their 95% pointwise confidence intervals for $i = 1,2,3$ and $j = 0,1,2,3$. The three columns present the estimated coefficients for areas one to three respectively (with subscript i); while the four rows depict the estimated coefficients for the intercept term, the coal, the crude oil, and the natural gas, respectively (with subscript j). The blue dash-dotted lines are the estimated varying-coefficient $\alpha_{ij}(U)$'s; the red dashed lines are the corresponding 95% pointwise confidence intervals.

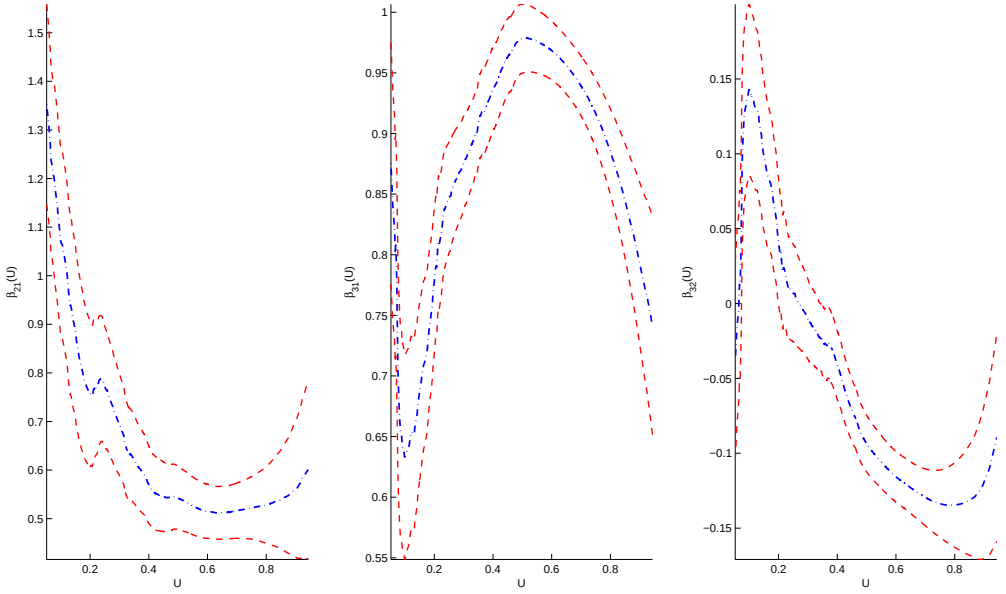

Figure 4.10. Estimated smoothing functions $\beta_{kj}(U)$'s and their 95% pointwise confidence intervals for $k = 2,3$ and $j = 1,\ldots,k-1$. The blue dash-dotted lines are the estimated $\beta_{kj}(U)$'s; the red dashed lines are the corresponding 95% pointwise confidence intervals.

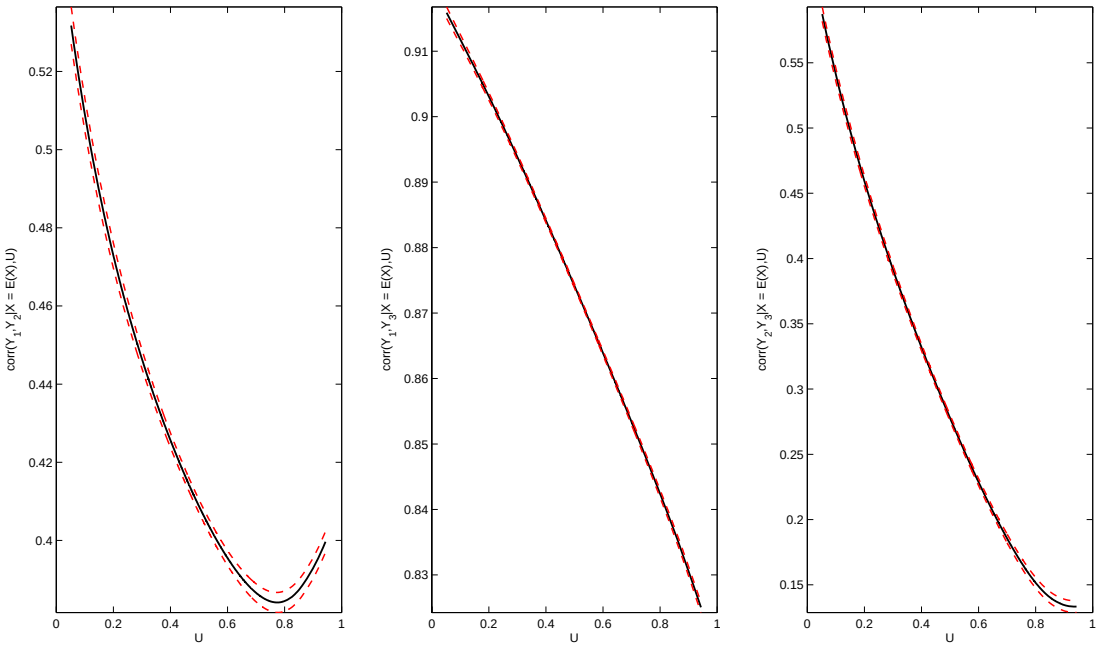

Figure 4.11. Estimated conditional correlation functions and their 95% pointwise confidence intervals when the covariate vector $\mathbf{x}$ is fixed at its sample average $\bar{\mathbf{x}}$. The black solid lines are the estimated conditional correlation functions; the red dashed lines are the corresponding 95% pointwise confidence intervals.

4.4 Technical Proofs

In this section, we will first provide three lemmas and then use these lemmas repetitively to prove the theorems. The following notation will be used in the proof of the theorems.

Lemma 4.4.1 (Li et al. (2012))**.** *Let* $\{(U_i,Y_i), i=1,\ldots,n\}$ *be independent and identically distributed random vectors. In addition to conditions (C1) and (C2), we further assume that* $\mathbb{E}(Y^2)<\infty$ *and* $\int y^2 f_{u,y}(u,y)dy<\infty$ *where* $f_{u,y}(u,y)$ *denotes the joint density of* (U,Y)*. Then*

$$\begin{aligned}&\sup_{U_0\in supp(U)}\Big|n^{-1}\sum_{i=1}^{n}K_h(U_i-U_0)Y_i\{(U_i-U_0)/h\}^l\\&\quad-\mathbb{E}\left[K_h(U_i-U_0)Y_i\{(U_i-U_0)/h\}^l\right]\Big|\\=\;& O_p\{\log n/(nh)^{1/2}\}, a.s.\end{aligned}\tag{4.38}$$

Lemma 4.4.2 (Li et al. (2012))**.** *Let* $\{(U_i,Y_i), i=1,\ldots,n\}$ *be independent and identically distributed random vectors. In addition to conditions (C1) and (C2), we further assume that* $g(u)=\int y f_{u,y}(u,y)dy<\infty$ *has a continuous second-order derivative, where* $f_{u,y}(u,y)$ *denotes the joint density of* (U,Y)*. Then*

$$\begin{aligned}&\sup_{U_0\in supp(U)}\left|\mathbb{E}\left[K_h(U_i-U_0)Y_i\{(U_i-U_0)/h\}^l\right]-\mu_l g(U_0)-h\mu_{l+1}\dot{g}(U_0)\right|\\=\;& O(h^2), a.s.\end{aligned}$$

Lemma 4.4.3 (Li et al. (2012))**.** *Suppose conditions (C1) - (C3) hold true. Then,*

$$\begin{aligned}&\sup_{U_0\in supp(U)}|\hat{\boldsymbol{\alpha}}_k(U_0)-\boldsymbol{\alpha}_k(U_0)|=O_p\{c_n(h_n)\},\ a.s.\\&\hat{\boldsymbol{\alpha}}_k(U_0)-\boldsymbol{\alpha}_k(U_0)=h_n^2\mu_2\ddot{\boldsymbol{\alpha}}_k(U_0)/2+\\=\;&\mathbf{C}_0\left\{\begin{pmatrix}1&0\\0&\mu_2\end{pmatrix}\otimes\mathbb{E}(\mathbf{x}\mathbf{x}^T|U=U_0)f(U_0)\right\}^{-1}\{\mathbf{B}^T(U_0)\mathbf{W}(U_0)\boldsymbol{\varepsilon}_k\}+o_p\{c_n(h_n)\},\end{aligned}$$

where $\mathbf{C}_0=(\mathbf{I}_d,\mathbf{0}_d)$, $\boldsymbol{\varepsilon}_k=(\varepsilon_{1k},\ldots,\varepsilon_{nk})^T$, $\mathbf{W}(U_0)=diag\{K_{h_n}(U_1-U_0),\ldots,K_{h_n}(U_n-$

$U_0)\}$, *and*

$$\mathbf{B}(U_0) = \begin{bmatrix} \mathbf{x}_1^T & (U_1 - U_0)\mathbf{x}_1^T/h_n \\ \mathbf{x}_2^T & (U_2 - U_0)\mathbf{x}_2^T/h_n \\ \vdots & \vdots \\ \mathbf{x}_n^T & (U_n - U_0)\mathbf{x}_n^T/h_n \end{bmatrix}.$$

Proof of Theorem 4.2.1 . Denote by $\mathbf{K}(U_0) = \text{diag}\{K_{h_1}(U_1 - U_0), \ldots, K_{h_1}(U_n - U_0)\}$, and

$$\hat{\mathbf{D}}(U_0) = \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T & (U_1 - U_0)\hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T/h_1 \\ \hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T & (U_2 - U_0)\hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T/h_1 \\ \vdots & \vdots \\ \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T & (U_n - U_0)\hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T/h_1 \end{bmatrix}.$$

For simplicity, we shortly denote

$$\mathbf{K}_0 = \mathbf{K}(U_0), \quad \hat{\mathbf{D}}_0 = \hat{\mathbf{D}}(U_0), \quad \Delta_i = U_i - U_0,\ i = 1, 2, \cdots, n.$$

As noted before, the local linear regression is to find $\{(\mathbf{a}_k, \mathbf{b}_k), k = 2, \ldots, m\}$ minimizing the least square function

$$\ell_k(\mathbf{a}_k, \mathbf{b}_k) = \sum_{i=1}^{n} \left[\varepsilon_{ik}^* - \sum_{j=1}^{k-1} \left\{ a_{kj} + b_{kj}(U_i - U_0) \right\} \varepsilon_{ij} \right]^2 K_h(U_i - U_0),$$

of which an optimal solutions is

$$\{\hat{\mathbf{a}}_k, h_1\hat{\mathbf{b}}_k\}^T = \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\boldsymbol{\varepsilon}}_k^*\} \tag{4.39}$$

Therefore,

$$\begin{aligned} &\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_0), h_1\hat{\dot{\boldsymbol{\beta}}}_{k,(k-1)}(U_0)\}^T \\ &= \{\hat{\mathbf{a}}_k, h_1\hat{\mathbf{b}}_k\}^T \\ &= \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\boldsymbol{\varepsilon}}_k^*\} \\ &= \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\boldsymbol{\varepsilon}_k^*\} + \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0(\hat{\boldsymbol{\varepsilon}}_k^* - \boldsymbol{\varepsilon}_k^*)\}, \end{aligned} \tag{4.40}$$

where $\hat{\boldsymbol{\varepsilon}}_k^* = (\hat{\varepsilon}_{1k}^*, \ldots, \hat{\varepsilon}_{nk}^*)^T$ and $\boldsymbol{\varepsilon}_k^* = (\varepsilon_{1k}^*, \ldots, \varepsilon_{nk}^*)^T$.

By (4.7) and Taylor expansion, we obtain that

$$\begin{aligned}\varepsilon_{ik}^* &= \hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T\{\boldsymbol{\beta}_{k,(k-1)}(U_0) + \Delta_i\dot{\boldsymbol{\beta}}_{k,(k-1)}(U_0) + \frac{\Delta_i^2}{2}\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_i^*)\} \\ &\quad + \{(\boldsymbol{\varepsilon}_{i(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T)\}\boldsymbol{\beta}_{k,(k-1)}(U_i) + e_{ik},\end{aligned}$$

where U_i^* is a value between U_0 and U_i. Thus,

$$\begin{aligned}&\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\boldsymbol{\varepsilon}_k^*\} \\ &= \{\boldsymbol{\beta}_{k,(k-1)}(U_0), h_1\dot{\boldsymbol{\beta}}_{k,(k-1)}(U_0)\}^T \\ &\quad + \frac{1}{2}\{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\hat{\mathbf{D}}_0^T\mathbf{K}_0\begin{bmatrix}\Delta_1^2\hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_1^*) \\ \Delta_2^2\hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_2^*) \\ \vdots \\ \Delta_n^2\hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_n^*)\end{bmatrix} \\ &\quad + \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\hat{\mathbf{D}}_0^T\mathbf{K}_0\begin{bmatrix}(\boldsymbol{\varepsilon}_{1(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_1) \\ (\boldsymbol{\varepsilon}_{2(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_2) \\ \vdots \\ (\boldsymbol{\varepsilon}_{n(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_n)\end{bmatrix} \\ &\quad + \{\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0\}^{-1}\hat{\mathbf{D}}_0^T\mathbf{K}_0\mathbf{e}_k. \qquad (4.41)\end{aligned}$$

Partition $\hat{\mathbf{D}}_0^T\mathbf{K}_0\hat{\mathbf{D}}_0$ in block form as

$$\begin{bmatrix}\mathbf{S}_{n0}(U_0) & \mathbf{S}_{n1}(U_0) \\ \mathbf{S}_{n1}(U_0) & \mathbf{S}_{n2}(U_0)\end{bmatrix}$$

where $\mathbf{S}_{nl}(U_0) = \sum_{i=1}^n K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\hat{\boldsymbol{\varepsilon}}_{i(k-1)}\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T$. In addition, $\mathbf{S}_{nl}(U_0)$ can be written as follows.

$$\begin{aligned}&\mathbf{S}_{nl}(U_0) \\ &= \sum_{i=1}^n K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\{\boldsymbol{\varepsilon}_{i(k-1)} + \hat{\boldsymbol{\varepsilon}}_{i(k-1)} - \boldsymbol{\varepsilon}_{i(k-1)}\}\{\boldsymbol{\varepsilon}_{i(k-1)} + \hat{\boldsymbol{\varepsilon}}_{i(k-1)} - \boldsymbol{\varepsilon}_{i(k-1)}\}^T \\ &= \sum_{i=1}^n K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\boldsymbol{\varepsilon}_{i(k-1)}\boldsymbol{\varepsilon}_{i(k-1)}^T\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{n} K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}\boldsymbol{\varepsilon}_{i(k-1)}^T\\
&+\sum_{i=1}^{n} K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\boldsymbol{\varepsilon}_{i(k-1)}\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}^T\\
&+\sum_{i=1}^{n} K_{h_1}(\Delta_i)\{\Delta_i/h_1\}^l\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}^T\\
=\;& \mathbf{S}_{nl,1}+\mathbf{S}_{nl,2}+\mathbf{S}_{nl,3}+\mathbf{S}_{nl,4}. \qquad (4.42)
\end{aligned}
$$

In the sequel, we deal with $\mathbf{S}_{nl,i}$ separately for $i=1,\ldots,4$.

First, let us study the convergence rate of $\mathbf{S}_{nl,1}$. Let $\mathbf{D}_1(U_0)=\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|U=U_0)$. By Lemma 4.4.1 and 4.4.2, we have that

$$
\begin{aligned}
\sup_{U_0\in\text{supp}\{U\}}|n^{-1}\mathbf{S}_{nl,1}(U_0)-\mathbf{D}_1(U_0)f(U_0)\mu_l| &= O_p\{h_1^2+\log n/(nh_1)^{1/2}\}\\
&= O_p\{c_n(h_1)\},\quad l=0,2\\
\sup_{U_0\in\text{supp}\{U\}}|n^{-1}\mathbf{S}_{n1,1}(U_0)| &= O_p\{h_1+\log n/(nh_1)^{1/2}\}. \qquad (4.43)
\end{aligned}
$$

Next, we study the convergence rate of $\mathbf{S}_{nl,i}, i=2,3,4$. Because these three terms are similar, we will only present the details of the convergence rate of $\mathbf{S}_{nl,2}$ for simplicity.

Recall that

$$
\boldsymbol{\varepsilon}_{i(k-1)}=\mathbf{y}_{i(k-1)}-(\boldsymbol{\alpha}_1(U_i),\ldots,\boldsymbol{\alpha}_{k-1}(U_i))^T\mathbf{x}_i=\mathbf{y}_{i(k-1)}-(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\boldsymbol{\alpha}_{(k-1)}(U_i),
$$

where $\boldsymbol{\alpha}_{(k-1)}(U_i)=(\boldsymbol{\alpha}_1^T(U_i),\ldots,\boldsymbol{\alpha}_{(k-1)}^T(U_i))^T$. Then,

$$
\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}=(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\boldsymbol{\alpha}_{(k-1)}(U_i)-\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)\}.
$$

Therefore, it can be obtained that

$$
\begin{aligned}
\mathbf{S}_{nl,2} &= \sum_{i=1}^{n} K_{h_1}\left(\frac{\Delta_i}{h_1}\right)^l(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\boldsymbol{\alpha}_{(k-1)}(U_i)-\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)\}\boldsymbol{\varepsilon}_{i(k-1)}^T\\
&= \sum_{i=1}^{n} K_{h_1}\left(\frac{\Delta_i}{h_1}\right)^l\{\boldsymbol{\alpha}_{(k-1)}^T(U_i)-\hat{\boldsymbol{\alpha}}_{(k-1)}^T(U_i)\}(\mathbf{I}_{k-1}\otimes\mathbf{x}_i)\boldsymbol{\varepsilon}_{i(k-1)}^T.
\end{aligned}
$$

In addition, let

$$\mathbf{S}^*_{nl,2} = \sum_{i=1}^{n} K_{h_1}(\Delta_i) \left(\frac{\Delta_i}{h_1}\right)^l (\mathbf{I}_{k-1} \otimes \mathbf{x}_i)\boldsymbol{\varepsilon}^T_{i(k-1)}.$$

By noting that $\mathbb{E}(\boldsymbol{\varepsilon}_{i(k-1)}|\mathbf{x}_i, U_i) = \mathbf{0}$, we have

$$\mathbb{E}(\mathbf{S}^*_{nl,2}) = n\mathbb{E}\left\{K_{h_1}(\Delta_i) \left(\frac{\Delta_i}{h_1}\right)^l (\mathbf{I}_{k-1} \otimes \mathbf{x}_i)\boldsymbol{\varepsilon}^T_{i(k-1)}\right\} = \mathbf{0}.$$

Additionally,

$$\begin{aligned}
\mathrm{Var}(\mathbf{S}^*_{nl,2}) &= \mathbb{E}(\mathbf{S}^{*2}_{nl,2}) - \{\mathbb{E}(\mathbf{S}^*_{nl,2})\}^2 \\
&= \mathbb{E}(\mathbf{S}^{*2}_{nl,2}) \\
&= n\mathbb{E}\left[K^2_{h_1}(\Delta_i) \left(\frac{\Delta_i}{h_1}\right)^{2l} \left\{(\mathbf{I}_{k-1} \otimes \mathbf{x}_i)\boldsymbol{\varepsilon}^T_{i(k-1)}\right\}^2\right] \\
&= O(n/h_1).
\end{aligned}$$

Therefore, by using Lemma 4.4.1 and Lemma 4.4.2 , we can obtain that

$$\sup_{U_0 \in \mathrm{supp}\{U\}} |n^{-1}\mathbf{S}^*_{nl,2}| = O_p\{\log n/(nh_1)^{1/2}\}.$$

Moreover, it follows from Lemma 4.4.3 that

$$\sup_{U_0 \in \mathrm{supp}\{U\}} |\boldsymbol{\alpha}^T_{(k-1)}(U_0) - \hat{\boldsymbol{\alpha}}^T_{(k-1)}(U_0)| = O_p\{c_n(h_n)\}.$$

Then,

$$\sup_{U_0 \in \mathrm{supp}\{U\}} |n^{-1}\mathbf{S}_{nl,2}| = O_p\left[c_n(h_n)\{\log n/(nh_1)^{1/2}\}\right], \tag{4.44}$$

which is asymptotically ignorable comparing with $n^{-1}\mathbf{S}_{nl,1}$. Similarly, it can be obtained that

$$\sup_{U_0 \in \mathrm{supp}\{U\}} |n^{-1}\mathbf{S}_{nl,3}| = O_p\left[c_n(h_n)\{\log n/(nh_1)^{1/2}\}\right], \tag{4.45}$$

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1}\mathbf{S}_{nl,4}| \;=\; O_p\{c_n^2(h_n)\}, \tag{4.46}$$

both of which are asymptotically ignorable comparing with $n^{-1}\mathbf{S}_{nl,1}$. Accordingly, it follows that

$$\begin{aligned} &\sup_{U_0 \in \text{supp}\{U\}} \left| n^{-1}\hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{D}}_0 - \begin{bmatrix} 1 & 0 \\ 0 & \mu_2 \end{bmatrix} \otimes \mathbf{D}_1(U_0) f(U_0) \right| \\ &= \begin{bmatrix} O_p\{c_n(h_1)\} & O_p\{h_1 + \log n/(nh_1)^{1/2}\} \\ O_p\{h_1 + \log n/(nh_1)^{1/2}\} & O_p\{c_n(h_1)\} \end{bmatrix} \end{aligned} \tag{4.47}$$

Next, let us study the convergence of the second term in (4.41).

$$\begin{aligned} &n^{-1}\hat{\mathbf{D}}_0^T \mathbf{K}_0 \begin{bmatrix} \Delta_1^2 \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_1^*) \\ \Delta_2^2 \hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_2^*) \\ \vdots \\ \Delta_n^2 \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_n^*) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i^2 \hat{\varepsilon}_{1(k-1)} \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_i^*) \\ \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i^3/h_1 \hat{\varepsilon}_{1(k-1)} \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_i^*) \end{bmatrix}. \end{aligned} \tag{4.48}$$

It follows from condition (C3) and the mean-value theorem that

$$|\ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_i^*) - \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_0)| \le c h_1 |\Delta_i|.$$

Therefore,

$$\begin{aligned} &\sup_{U_0 \in \text{supp}\{U\}} \left| n^{-1}\hat{\mathbf{D}}_0^T \mathbf{K}_0 \begin{bmatrix} \Delta_1^2 \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_1^*)/2 \\ \Delta_2^2 \hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_2^*)/2 \\ \vdots \\ \Delta_n^2 \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_n^*)/2 \end{bmatrix} \right. \\ &\qquad \left. -\frac{h_1^2}{2} f(U_0) \begin{bmatrix} \mu_2 \\ 0 \end{bmatrix} \mathbb{E}\{\boldsymbol{\varepsilon}_{(k-1)} \boldsymbol{\varepsilon}_{(k-1)}^T | U = U_0\} \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_0) \right| \\ &= \begin{bmatrix} O_P\{h_1^2(h_1^2 + \log n/(nh_1)^{1/2})\} \\ O_P\{h_1^2(h_1 + \log n/(nh_1)^{1/2})\} \end{bmatrix}. \end{aligned} \tag{4.49}$$

It remains to study the convergence of the third term in (4.41). It is noticed that the elements in the matrix $\mathbf{D}$ have mean zero asymptotically. Similarly using $\hat{\boldsymbol{\varepsilon}}_{i(k-1)} - \boldsymbol{\varepsilon}_{i(k-1)} = (\mathbf{I}_{k-1} \otimes \mathbf{x}_i^T)\{\boldsymbol{\alpha}_{(k-1)}(U_i) - \hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)\}$ and the uniform convergence results obtained in Lemma 3, we have

$$\sup_{U_0 \in \text{supp}\{U\}} \left| \hat{\mathbf{D}}_0^T \mathbf{K}_0 \begin{bmatrix} (\boldsymbol{\varepsilon}_{1(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_1) \\ (\boldsymbol{\varepsilon}_{2(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{2(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_2) \\ \vdots \\ (\boldsymbol{\varepsilon}_{n(k-1)}^T - \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T)\boldsymbol{\beta}_{k,(k-1)}(U_n) \end{bmatrix} \right| = o_P\{c_n(h_1)\}. \tag{4.50}$$

Therefore, by combining the above results, we can have that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{k,(k-1)}(U_0) - \boldsymbol{\beta}_{k,(k-1)}(U_0) &= \frac{1}{2}\mu_2 h_1^2 \ddot{\boldsymbol{\beta}}_{k,(k-1)}(U_0) + \\ \mathbf{C}\left\{ \begin{bmatrix} 1 & 0 \\ 0 & \mu_2 \end{bmatrix} \otimes \mathbf{D}_1(U_0) f(U_0) \right\}^{-1} &\{\mathbf{D}^T(U_0)\mathbf{K}(U_0)\mathbf{e}_k\} + o_P\{c_n(h_1)\}, \end{aligned} \tag{4.51}$$

where $\mathbf{C} = [\mathbf{I}_{d(k-1)\times d(k-1)}, \mathbf{0}_{d(k-1)\times d(k-1)}]$. Then the asymptotic normality can be easily derived. The proof of this theorem is thus completed.

□

Proof of Theorem 4.2.2 . For simplicity, we shortly denote

$$\mathbf{K}_i = \mathbf{K}(U_i), \quad \hat{\mathbf{D}}_i = \hat{\mathbf{D}}(U_i), \quad \Delta_i = U_i - U_0,\ i = 1, 2, \cdots, n.$$

We have that

$$\hat{\boldsymbol{\gamma}}_k = \left\{\hat{\mathbf{V}}_k^T (\mathbf{I} - \hat{\mathbf{S}}_{h_1})^T (\mathbf{I} - \hat{\mathbf{S}}_{h_1}) \hat{\mathbf{V}}_k\right\}^{-1} \left\{\hat{\mathbf{V}}_k^T (\mathbf{I} - \hat{\mathbf{S}}_{h_1})^T (\mathbf{I} - \hat{\mathbf{S}}_{h_1}) \hat{\boldsymbol{\varepsilon}}_k\right\}, \tag{4.52}$$

where $\hat{\mathbf{V}}_k = (\hat{\mathbf{v}}_{1k}, \ldots, \hat{\mathbf{v}}_{nk})^T = (\hat{\boldsymbol{\varepsilon}}_{1(k-1)} \otimes \mathbf{x}_1, \ldots, \hat{\boldsymbol{\varepsilon}}_{n(k-1)} \otimes \mathbf{x}_n)^T$,

$$\hat{\mathbf{S}}_{h_1} = \begin{bmatrix} (\hat{\boldsymbol{\varepsilon}}_{1(k-1)}, \mathbf{0})\left\{\hat{\mathbf{D}}_1^T \mathbf{K}_1 \hat{\mathbf{D}}_1\right\}^{-1} \hat{\mathbf{D}}_1^T \mathbf{K}_1 \\ \vdots \\ (\hat{\boldsymbol{\varepsilon}}_{n(k-1)}, \mathbf{0})\left\{\hat{\mathbf{D}}_n^T \mathbf{K}_n \hat{\mathbf{D}}_n\right\}^{-1} \hat{\mathbf{D}}_n^T \mathbf{K}_n \end{bmatrix}. \tag{4.53}$$

Denote $\boldsymbol{\delta}_{ik} = \hat{\boldsymbol{\varepsilon}}_{ik} - \boldsymbol{\varepsilon}_{ik}$ and $\boldsymbol{\delta}_{i(k-1)} = \hat{\boldsymbol{\varepsilon}}_{i(k-1)} - \boldsymbol{\varepsilon}_{i(k-1)}$. Then $\hat{\mathbf{D}}_0$ can be re-written as

$$\begin{aligned}\hat{\mathbf{D}}_0 &= \begin{bmatrix} \boldsymbol{\delta}_{1(k-1)}^T & \Delta_1 \boldsymbol{\delta}_{1(k-1)}^T/h_1 \\ \boldsymbol{\delta}_{2(k-1)}^T & \Delta_2 \boldsymbol{\delta}_{2(k-1)}^T/h_1 \\ \vdots & \vdots \\ \boldsymbol{\delta}_{n(k-1)}^T & \Delta_n \boldsymbol{\delta}_{n(k-1)}^T/h_1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1(k-1)}^T & \Delta_1 \boldsymbol{\varepsilon}_{1(k-1)}^T/h_1 \\ \boldsymbol{\varepsilon}_{2(k-1)}^T & \Delta_2 \boldsymbol{\varepsilon}_{2(k-1)}^T/h_1 \\ \vdots & \vdots \\ \boldsymbol{\varepsilon}_{n(k-1)}^T & \Delta_n \boldsymbol{\varepsilon}_{n(k-1)}^T/h_1 \end{bmatrix} \\ &= \boldsymbol{\delta}_{\mathbf{D}_0} + \mathbf{D}_0.\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{D}}_0 &= (\boldsymbol{\delta}_{\mathbf{D}_0} + \mathbf{D}_0)^T \mathbf{K}_0 (\boldsymbol{\delta}_{\mathbf{D}_0} + \mathbf{D}_0) \\ &= \mathbf{D}_0^T \mathbf{K}_0 \mathbf{D}_0 + \boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0 \\ &\quad + \mathbf{D}_0^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{D}_0} + \boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{D}_0}. \end{aligned} \tag{4.54}$$

We first study the convergence rate of $\boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0$. We have that

$$\begin{aligned}&\boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0 \\ &= \begin{bmatrix} \sum_{i=1}^n K_{h_1}(\Delta_i) \boldsymbol{\varepsilon}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T & \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i/h_1 \boldsymbol{\varepsilon}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T \\ \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i/h_1 \boldsymbol{\varepsilon}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T & \sum_{i=1}^n K_{h_1}(\Delta_i) (\Delta_i/h_1)^2 \boldsymbol{\varepsilon}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T \end{bmatrix}.\end{aligned}$$

It is noted that each block in the matrix $\boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0$ has the similar form as that of $\mathbf{S}_{nl,3}$. Therefore, it can be obtained that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0| = O_p\left[c_n(h_n)\{\log n/(nh_1)^{1/2}\}\right] \tag{4.55}$$

Next, let us study the convergence rate of $\boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{D}_0}$, which can be re-formulated as follows:

$$\begin{aligned}&\boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{D}_0 \\ &= \begin{bmatrix} \sum_{i=1}^n K_{h_1}(\Delta_i) \boldsymbol{\delta}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T & \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i/h_1 \boldsymbol{\delta}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T \\ \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i/h_1 \boldsymbol{\delta}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T & \sum_{i=1}^n K_{h_1}(\Delta_i) (\Delta_i/h_1)^2 \boldsymbol{\delta}_{i(k-1)} \boldsymbol{\delta}_{i(k-1)}^T \end{bmatrix}.\end{aligned}$$

Based on $\mathbf{S}_{nl,4}$ and its corresponding convergence rate, we have that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{D}_0}| \;=\; O_p\left\{c_n^2(h_n)\right\} \tag{4.56}$$

Therefore, by combining the results of (4.54),(4.55), and (4.56), we can have that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{D}}_0 - n^{-1} \mathbf{D}_0^T \mathbf{K}_0 \mathbf{D}_0| \;=\; O_p\left\{c_n^2(h_n)\right\} \tag{4.57}$$

Next, let us derive the similar property for the term $\hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{V}}_k$. We have that

$$\begin{aligned}
\hat{\mathbf{V}}_k &= \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \otimes \mathbf{x}_1^T \\ \vdots \\ \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T \otimes \mathbf{x}_n^T \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\varepsilon}_{1(k-1)}^T \otimes \mathbf{x}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_{n(k-1)}^T \otimes \mathbf{x}_n^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{\delta}_{1(k-1)}^T \otimes \mathbf{x}_1^T \\ \vdots \\ \boldsymbol{\delta}_{n(k-1)}^T \otimes \mathbf{x}_n^T \end{bmatrix} \\
&= \mathbf{V}_k + \boldsymbol{\delta}_{\mathbf{V}_k}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{V}}_k &= (\boldsymbol{\delta}_{\mathbf{D}_0} + \mathbf{D}_0)^T \mathbf{K}_0 (\mathbf{V}_k + \boldsymbol{\delta}_{\mathbf{V}_k}) \\
&= \mathbf{D}_0^T \mathbf{K}_0 \mathbf{V}_k + \boldsymbol{\delta}_{\mathbf{D}_0}^T \mathbf{K}_0 \mathbf{V}_k \\
&\quad + \mathbf{D}_0^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{V}_k} + \boldsymbol{\delta}_{\mathbf{V}_k}^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{V}_k}.
\end{aligned} \tag{4.58}$$

It is trivial to derive that

$$\begin{aligned}
&\boldsymbol{\delta}_{\mathbf{V}_k}^T \mathbf{K}_0 \mathbf{D}_0 \\
&= \begin{bmatrix} \sum_{i=1}^n K_{h_1}(\Delta_i) \boldsymbol{\delta}_{i(k-1)} (\boldsymbol{\varepsilon}_{i(k-1)}^T \otimes \mathbf{x}_i^T) \\ \sum_{i=1}^n K_{h_1}(\Delta_i) \Delta_i / h_1 \boldsymbol{\delta}_{i(k-1)} (\boldsymbol{\varepsilon}_{i(k-1)}^T \otimes \mathbf{x}_i^T) \end{bmatrix}.
\end{aligned}$$

By using the same argument that leads to $\mathbf{S}_{nl,3}$ and its corresponding convergence rate,

we have that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \boldsymbol{\delta}_{\mathbf{V}_k}^T \mathbf{K}_0 \mathbf{D}_0| = O_p \left[c_n(h_n) \{ \log n / (nh_1)^{1/2} \} \right] \tag{4.59}$$

Accordingly, it follow from $\mathbf{S}_{nl,4}$ that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \boldsymbol{\delta}_{\mathbf{V}_k}^T \mathbf{K}_0 \boldsymbol{\delta}_{\mathbf{V}_k}| = O_p \left\{ c_n^2(h_n) \right\} \tag{4.60}$$

Thus, combining the results of (4.59) and (4.60), we can obtain that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{V}}_k - n^{-1} \mathbf{D}_0^T \mathbf{K}_0 \mathbf{V}_k| \quad = \quad O_p \left[c_n(h_n) \left\{ \frac{\log n}{(nh_1)^{1/2}} + c_n(h_n) \right\} \right] \tag{4.61}$$

Denote $\hat{\mathbf{Q}}_0 = (\hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{D}}_0)^{-1} \hat{\mathbf{D}}_0^T \mathbf{K}_0 \hat{\mathbf{V}}_k$ and $\mathbf{Q}_0 = (\mathbf{D}_0^T \mathbf{K}_0 \mathbf{D}_0)^{-1} \mathbf{D}_0^T \mathbf{K}_0 \mathbf{V}_k$, then combining the results of (4.57) and (4.61) yields that

$$\sup_{U_0 \in \text{supp}\{U\}} |n^{-1} \hat{\mathbf{Q}}_0 - n^{-1} \mathbf{Q}_0| \quad = \quad O_p \left[c_n(h_n) \left\{ \frac{\log n}{(nh_1)^{1/2}} + c_n(h_n) \right\} \right]. \tag{4.62}$$

Now, let's consider the term $(\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k)^T (\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k)$. Note that

$$\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k = \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T \otimes \mathbf{x}_1^T \\ \vdots \\ \hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T \otimes \mathbf{x}_n^T \end{bmatrix} - \begin{bmatrix} (\hat{\boldsymbol{\varepsilon}}_{1(k-1)}^T, \mathbf{0}) \hat{\mathbf{Q}}_1 \\ \vdots \\ (\hat{\boldsymbol{\varepsilon}}_{n(k-1)}^T, \mathbf{0}) \hat{\mathbf{Q}}_n \end{bmatrix}. \tag{4.63}$$

Therefore, $(\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k)^T (\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k)$ can be decomposed as

$$\begin{aligned}
&(\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k)^T (\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1} \hat{\mathbf{V}}_k) \\
&= \sum_{i=1}^{n} \left[\left\{ \hat{\boldsymbol{\varepsilon}}_{i(k-1)} \otimes \mathbf{x}_i - \hat{\mathbf{Q}}_i^T (\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T, \mathbf{0})^T \right\} \left\{ \hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T \otimes \mathbf{x}_i^T - (\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T, \mathbf{0}) \hat{\mathbf{Q}}_i \right\} \right] \\
&= \sum_{i=1}^{n} \left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)} \otimes \mathbf{x}_i \right) \left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T \otimes \mathbf{x}_i^T \right) \\
&\quad - \sum_{i=1}^{n} \left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)} \otimes \mathbf{x}_i \right) \left((\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T, \mathbf{0}) \hat{\mathbf{Q}}_i \right)
\end{aligned}$$

$$
\begin{aligned}
&-\sum_{i=1}^{n}\left(\hat{\mathbf{Q}}_i^T(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T,\mathbf{0})^T\right)\left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right)\\
&+\sum_{i=1}^{n}\left(\hat{\mathbf{Q}}_i^T(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T,\mathbf{0})^T\right)\left((\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T,\mathbf{0})\hat{\mathbf{Q}}_i\right)\\
=\;&\hat{\mathbf{E}}_1+\hat{\mathbf{E}}_2+\hat{\mathbf{E}}_3+\hat{\mathbf{E}}_4. \qquad (4.64)
\end{aligned}
$$

In the sequel, we deal with $\hat{\mathbf{E}}_i$ separately for $i=1,\ldots,4$. We fist consider $\hat{\mathbf{E}}_1$.

$$
\begin{aligned}
\hat{\mathbf{E}}_1 &= \sum_{i=1}^{n}\left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}\otimes\mathbf{x}_i\right)\left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right)\\
&= \sum_{i=1}^{n}\left\{\left(\boldsymbol{\varepsilon}_{i(k-1)}+\boldsymbol{\delta}_{i(k-1)}\right)\otimes\mathbf{x}_i\right\}\left\{\left(\boldsymbol{\varepsilon}_{i(k-1)}^T+\boldsymbol{\delta}_{i(k-1)}^T\right)\otimes\mathbf{x}_i^T\right\}\\
&= \mathbf{E}_1+\boldsymbol{\delta}_{\mathbf{E}_1}, \qquad (4.65)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{E}_1 &= \sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i\right)\left(\boldsymbol{\varepsilon}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right) \qquad (4.66)\\
\boldsymbol{\delta}_{\mathbf{E}_1} &= \sum_{i=1}^{n}\left(\boldsymbol{\delta}_{i(k-1)}\otimes\mathbf{x}_i\right)\left(\boldsymbol{\varepsilon}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right)+\sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i\right)\left(\boldsymbol{\delta}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right)\\
&\quad+\sum_{i=1}^{n}\left(\boldsymbol{\delta}_{i(k-1)}\otimes\mathbf{x}_i\right)\left(\boldsymbol{\delta}_{i(k-1)}^T\otimes\mathbf{x}_i^T\right) \qquad (4.67)
\end{aligned}
$$

By using the same argument as before, we have

$$
\sup_{U_0\in\text{supp}\{U\}}|n^{-1}\hat{\mathbf{E}}_1-n^{-1}\mathbf{E}_1|=O_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}}+c_n(h_n)\right\}\right]. \qquad (4.68)
$$

Now we turn to study the convergence rate of $\hat{\mathbf{E}}_2$. Note that

$$
\begin{aligned}
\hat{\mathbf{E}}_2 &= \sum_{i=1}^{n}\left(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}\otimes\mathbf{x}_i\right)\left((\hat{\boldsymbol{\varepsilon}}_{i(k-1)}^T,\mathbf{0})\hat{\mathbf{Q}}_i\right)\\
&= \sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i+\boldsymbol{\delta}_{i(k-1)}\otimes\mathbf{x}_i\right)\left((\boldsymbol{\varepsilon}_{i(k-1)}^T,\mathbf{0})\hat{\mathbf{Q}}_i+(\boldsymbol{\delta}_{i(k-1)}^T,\mathbf{0})\hat{\mathbf{Q}}_i\right)\\
&= \hat{\mathbf{E}}_{2,1}+\hat{\mathbf{E}}_{2,2}+\hat{\mathbf{E}}_{2,3}+\hat{\mathbf{E}}_{2,4},
\end{aligned}
$$

where

$$\hat{\mathbf{E}}_{2,1} = \sum_{i=1}^{n} \left(\boldsymbol{\varepsilon}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})\hat{\mathbf{Q}}_i\right) \tag{4.69}$$

$$\hat{\mathbf{E}}_{2,2} = \sum_{i=1}^{n} \left(\boldsymbol{\varepsilon}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\delta}_{i(k-1)}^T, \mathbf{0})\hat{\mathbf{Q}}_i\right) \tag{4.70}$$

$$\hat{\mathbf{E}}_{2,3} = \sum_{i=1}^{n} \left(\boldsymbol{\delta}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})\hat{\mathbf{Q}}_i\right) \tag{4.71}$$

$$\hat{\mathbf{E}}_{2,4} = \sum_{i=1}^{n} \left(\boldsymbol{\delta}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\delta}_{i(k-1)}^T, \mathbf{0})\hat{\mathbf{Q}}_i\right) \tag{4.72}$$

By (4.62), we obtain that

$$\hat{\mathbf{E}}_{2,1} = \sum_{i=1}^{n} \left(\boldsymbol{\varepsilon}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})\mathbf{Q}_i\right) + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}} + c_n(h_n)\right\}\right]. \tag{4.73}$$

Following the same argument, we have that $\hat{\mathbf{E}}_{2,2}$, $\hat{\mathbf{E}}_{2,3}$, and $\hat{\mathbf{E}}_{2,4}$ are asymptotically negligible comparing with $\hat{\mathbf{E}}_{2,1}$. Hence, if denoting

$$\mathbf{E}_2 = \sum_{i=1}^{n} \left(\boldsymbol{\varepsilon}_{i(k-1)} \otimes \mathbf{x}_i\right) \left((\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})\mathbf{Q}_i\right), \tag{4.74}$$

then it can be shown that

$$\hat{\mathbf{E}}_2 = \mathbf{E}_2 + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}} + c_n(h_n)\right\}\right]. \tag{4.75}$$

Similar properties can be shown for $\hat{\mathbf{E}}_3$ and $\hat{\mathbf{E}}_4$. Therefore,

$$\begin{aligned}
&(\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1}\hat{\mathbf{V}}_k)^T(\hat{\mathbf{V}}_k - \hat{\mathbf{S}}_{h_1}\hat{\mathbf{V}}_k) \\
&= \sum_{i=1}^{n}\left[\left\{\boldsymbol{\varepsilon}_{i(k-1)} \otimes \mathbf{x}_i - \mathbf{Q}_i^T(\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})^T\right\}\left\{\boldsymbol{\varepsilon}_{i(k-1)}^T \otimes \mathbf{x}_i^T - (\boldsymbol{\varepsilon}_{i(k-1)}^T, \mathbf{0})\mathbf{Q}_i\right\}\right] \\
&\quad + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}} + c_n(h_n)\right\}\right] \\
&= (\mathbf{V}_k - \mathbf{S}_{h_1}\mathbf{V}_k)^T(\mathbf{V}_k - \mathbf{S}_{h_1}\mathbf{V}_k) + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}} + c_n(h_n)\right\}\right]
\end{aligned} \tag{4.76}$$

Equivalently, by using the same argument as before, we can show that

$$\begin{aligned}
&\hat{\mathbf{V}}_k^T(\mathbf{I}-\hat{\mathbf{S}}_{h_1})^T(\mathbf{I}-\hat{\mathbf{S}}_{h_1})\hat{\boldsymbol{\varepsilon}}_k \\
&= \mathbf{V}_k^T(\mathbf{I}-\mathbf{S}_{h_1})^T(\mathbf{I}-\mathbf{S}_{h_1})\boldsymbol{\varepsilon}_k + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}}+c_n(h_n)\right\}\right]. \qquad (4.77)
\end{aligned}$$

Combining the results of (4.76) and (4.77), we have

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}_k &= \left\{\mathbf{V}_k^T(\mathbf{I}-\mathbf{S}_{h_1})^T(\mathbf{I}-\mathbf{S}_{h_1})\mathbf{V}_k\right\}^{-1}\left\{\mathbf{V}_k^T(\mathbf{I}-\mathbf{S}_{h_1})^T(\mathbf{I}-\mathbf{S}_{h_1})\boldsymbol{\varepsilon}_k\right\} \\
&\quad + o_p\left[c_n(h_n)\left\{\frac{\log n}{(nh_1)^{1/2}}+c_n(h_n)\right\}\right]. \qquad (4.78)
\end{aligned}$$

Denote

$$\tilde{\boldsymbol{\gamma}}_k = \left\{\mathbf{V}_k^T(\mathbf{I}-\mathbf{S}_{h_1})^T(\mathbf{I}-\mathbf{S}_{h_1})\mathbf{V}_k\right\}^{-1}\left\{\mathbf{V}_k^T(\mathbf{I}-\mathbf{S}_{h_1})^T(\mathbf{I}-\mathbf{S}_{h_1})\boldsymbol{\varepsilon}_k\right\}, \qquad (4.79)$$

then by following Theorem 4.1 of Fan and Huang (2004), we can have the asymptotic normality of $\tilde{\boldsymbol{\gamma}}_k$ as follows,

$$\sqrt{n}(\tilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) \quad \to_D \quad \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma\right), \qquad (4.80)$$

for $k = 2,3,\ldots,m$, and $\boldsymbol{\Sigma}_\gamma$ is given as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_\gamma &= \sigma_k^2(\mathbf{x},U)\Big[\mathbb{E}\left\{(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})^T\right\} \\
&\quad -\mathbb{E}\left\{\mathbb{E}\left\{(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})\mathbf{x}^T|U\right\}\mathbb{E}(\mathbf{x}\mathbf{x}^T|U)^{-1}\mathbb{E}\left\{\mathbf{x}(\boldsymbol{\varepsilon}_{(k-1)}\otimes\mathbf{x})^T|U\right\}\right\}\Big]^{-1}.
\end{aligned}$$

This, together with (4.78), proves the results.

□

Proof of Theorem 4.2.3 . Note that

$$\begin{aligned}
\hat{\sigma}_k^2(U_0) &= \sum_{i=1}^n\{\hat{e}_{ik}^2\}K_{h_2}\Big/\sum_{i=1}^n K_{h_2}(\Delta_i) \\
&= \sum_{i=1}^n\{e_{ik}^2+2e_{ik}(\hat{e}_{ik}-e_{ik})+(\hat{e}_{ik}-e_{ik})^2\}K_{h_2}(\Delta_i)\Big/\sum_{i=1}^n K_{h_2}(\Delta_i).
\end{aligned}$$

First we need to prove that

$$\sup_{U_0}\left|\sum_{i=1}^{n} e_{ik}(\hat{e}_{ik}-e_{ik})K_{h_2}(\Delta_i)\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\right| = O_p\left\{c_n(h_2)c_n(h_n)+c_n(h_1)\right\}, \quad (4.81)$$

$$\sup_{U_0}\left|\sum_{i=1}^{n} (\hat{e}_{ik}-e_{ik})^2K_{h_2}(\Delta_i)\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\right| = O_p\{c_n(h_2)c_n^2(h_n)+c_n^2(h_1)\}. \quad (4.82)$$

Using the fact that $\boldsymbol{\varepsilon}_{i(k-1)} = \mathbf{y}_{i(k-1)} - (\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\text{vec}\{\boldsymbol{\alpha}_{(k-1)}(U_i)\}$ and $\hat{\varepsilon}_{ik} = y_{ik} - \hat{\boldsymbol{\alpha}}_k^T(U_i)\mathbf{x}_i$, it can be derived that

$$\begin{aligned}
&\hat{e}_{ik}-e_{ik}\\
&= \mathbf{x}_i^T\{\boldsymbol{\alpha}_k(U_i)-\hat{\boldsymbol{\alpha}}_k(U_i)\} - \hat{\boldsymbol{\gamma}}_k^T(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}\otimes\mathbf{x}_i) + \boldsymbol{\gamma}_k^T(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i)\\
&\quad -\hat{\boldsymbol{\beta}}_{k,(k-1)}^T(U_i)\hat{\boldsymbol{\varepsilon}}_{i(k-1)} + \boldsymbol{\beta}_{k,(k-1)}^T(U_i)\boldsymbol{\varepsilon}_{i(k-1)}\\
&= \mathbf{x}_i^T\{\boldsymbol{\alpha}_k(U_i)-\hat{\boldsymbol{\alpha}}_k(U_i)\} - (\hat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^T(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i)\\
&\quad -(\hat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^T\{(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)})\otimes\mathbf{x}_i\} - \boldsymbol{\gamma}_k^T\{(\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)})\otimes\mathbf{x}_i\}\\
&\quad -\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_i)-\boldsymbol{\beta}_{k,(k-1)}(U_i)\}^T\boldsymbol{\varepsilon}_{i(k-1)}\\
&\quad -\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_i)-\boldsymbol{\beta}_{k,(k-1)}(U_i)\}^T\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}\\
&\quad -\boldsymbol{\beta}_{k,(k-1)}^T(U_i)\{\hat{\boldsymbol{\varepsilon}}_{i(k-1)}-\boldsymbol{\varepsilon}_{i(k-1)}\}\\
&= \mathbf{x}_i^T\{\boldsymbol{\alpha}_k(U_i)-\hat{\boldsymbol{\alpha}}_k(U_i)\} - (\hat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^T(\boldsymbol{\varepsilon}_{i(k-1)}\otimes\mathbf{x}_i)\\
&\quad -(\hat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^T\left(\left[(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)-\boldsymbol{\alpha}_{(k-1)}(U_i)\}\right]\otimes\mathbf{x}_i\right)\\
&\quad -\boldsymbol{\gamma}_k^T\left(\left[(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)-\boldsymbol{\alpha}_{(k-1)}(U_i)\}\right]\otimes\mathbf{x}_i\right)\\
&\quad -\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_i)-\boldsymbol{\beta}_{k,(k-1)}(U_i)\}^T\boldsymbol{\varepsilon}_{i(k-1)}\\
&\quad -\{\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_i)-\boldsymbol{\beta}_{k,(k-1)}(U_i)\}^T\left(\left[(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)-\boldsymbol{\alpha}_{(k-1)}(U_i)\}\right]\otimes\mathbf{x}_i\right)\\
&\quad -\boldsymbol{\beta}_{k,(k-1)}^T(U_i)\left(\left[(\mathbf{I}_{k-1}\otimes\mathbf{x}_i^T)\{\hat{\boldsymbol{\alpha}}_{(k-1)}(U_i)-\boldsymbol{\alpha}_{(k-1)}(U_i)\}\right]\otimes\mathbf{x}_i\right) \quad (4.83)
\end{aligned}$$

Denote $g_1(\mathbf{x})$ as a function of $\mathbf{x}$. By using condition (C5), Lemma 4.4.1 and 4.4.2 , $\mathbb{E}(e_{ik}|\mathbf{x}_i,U_i)=0$ implies

$$\sup_{U_0}\left|\sum_{i=1}^{n} e_{ik}g_1(\mathbf{x}_i)K_{h_2}(\Delta_i)\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\right| = O_p\{c_n(h_2)\}. \quad (4.84)$$

Therefore, by employing the results that

$$\begin{aligned}\sup_{U_0}|\hat{\boldsymbol{\alpha}}_k(U_0)-\boldsymbol{\alpha}_k(U_0)| &= O_p\{c_n(h_n)\},\\ \sup_{U_0}\left|\hat{\boldsymbol{\beta}}_{k,(k-1)}(U_0)-\boldsymbol{\beta}_{k,(k-1)}(U_0)\right| &= O_p\{c_n(h_1)\},\end{aligned}$$

(4.81) can be proven after some tedious calculation.

Next, let $g_2(\mathbf{x},\boldsymbol{\varepsilon}_{(k-1)})$ denote the function of $(\mathbf{x},\boldsymbol{\varepsilon}_{(k-1)})$. Similarly, by employing condition (C5), Lemma 4.4.1 and 4.4.2, it can be obtained that

$$\sup_{U_0}\left|\sum_{i=1}^{n} g_2(\mathbf{x}_i,\boldsymbol{\varepsilon}_{(k-1)})K_{h_2}(\Delta_i)\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\right| = O_p(1). \tag{4.85}$$

Then (4.82) can be derived by using Cauchy-Schwarz inequality. Therefore, (4.81) and (4.82) are proven.

Next we will show that

$$\begin{aligned}&\sum_{i=1}^{n}K_{h_2}(\Delta_i)e_{ik}^2\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)-\sigma_k^2(U_0)\\ &= h_2^2\mu_2\left\{\dot{\sigma}_k^2(U_0)\frac{\dot{f}(U_0)}{f(U_0)}+\frac{\ddot{\sigma}_k^2(U_0)}{2}\right\}\\ &\quad +n^{-1}\sum_{i=1}^{n}K_{h_2}(\Delta_i)\{e_{ik}^2-\sigma_k^2(U_i)\}/f(U_0)\\ &\quad +o_p\{c_n(h_2)\}.\end{aligned} \tag{4.86}$$

To prove equation (4.86), we first re-write the left part as

$$\begin{aligned}&\sum_{i=1}^{n}K_{h_2}(\Delta_i)e_{ik}^2\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)-\sigma_k^2(U_0)\\ &= \sum_{i=1}^{n}K_{h_2}(\Delta_i)\{e_{ik}^2-\sigma_k^2(U_0)\}\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\\ &= \sum_{i=1}^{n}K_{h_2}(\Delta_i)\{e_{ik}^2-\sigma_k^2(U_i)+\sigma_k^2(U_i)-\sigma_k^2(U_0)\}\Big/\sum_{i=1}^{n}K_{h_2}(\Delta_i)\end{aligned}$$

$$
\begin{aligned}
= & \sum_{i=1}^{n} K_{h_2}(\Delta_i)\{e_{ik}^2 - \sigma_k^2(U_i)\} \Big/ \sum_{i=1}^{n} K_{h_2}(\Delta_i) \\
& + \sum_{i=1}^{n} K_{h_2}(\Delta_i)\{\sigma_k^2(U_i) - \sigma_k^2(U_0)\} \Big/ \sum_{i=1}^{n} K_{h_2}(\Delta_i) \qquad (4.87)
\end{aligned}
$$

It is noted that $f(U_0)$ is bounded away from 0 by condition (C1). Thus Lemma 4.4.1 and Lemma 4.4.2 imply that

$$
\sup_{U_0 \in \text{supp}\{U\}} \left| \hat{f}^{-1}(U_0) - f^{-1}(U_0) \right| = O_p\{c_n(h_2)\}, \text{ almost surely}, \qquad (4.88)
$$

where $\hat{f}(U_0) = n^{-1} \sum_{i=1}^{n} K_{h_2}(\Delta_i)$.

Applying (4.88), equation (4.87) can be expressed as

$$
\begin{aligned}
& \sum_{i=1}^{n} K_{h_2}(\Delta_i) e_{ik}^2 \Big/ \sum_{i=1}^{n} K_{h_2}(\Delta_i) - \sigma_k^2(U_0) \\
= & \; n^{-1} \sum_{i=1}^{n} K_{h_2}(\Delta_i)\{e_{ik}^2 - \sigma_k^2(U_i)\} \left[f^{-1}(U_0) + O_p\{c_n(h_2)\} \right] \\
& + n^{-1} \sum_{i=1}^{n} K_{h_2}(\Delta_i)\{\sigma_k^2(U_i) - \sigma_k^2(U_0)\} \left[f^{-1}(U_0) + O_p\{c_n(h_2)\} \right] \qquad (4.89)
\end{aligned}
$$

Denote the first and second term in the above equality as E_1 and E_2 respectively. Then E_1 is an average of independent and identically distributed random variables. E_2 can be approximated as follows.

By Taylor expansion,

$$
\sigma_k^2(U_i) \approx \sigma_k^2(U_0) + (U_i - U_0)\dot{\sigma}_k^2(U_0) + (U_i - U_0)^2 \ddot{\sigma}_k^2(U_i^*)/2, \qquad (4.90)
$$

where U_i^* is a value between U_i and U_0. Then $\sum_{i=1}^{n} K_{h_2}(\Delta_i)\{\sigma_k^2(U_i) - \sigma_k^2(U_0)\} \Big/ f(U_0)$ can be approximated by

$$
\begin{aligned}
& n^{-1} \sum_{i=1}^{n} K_{h_2}(\Delta_i)\{\sigma_k^2(U_i) - \sigma_k^2(U_0)\} \Big/ f(U_0) \\
\approx & \; \frac{\sum_{i=1}^{n} K_{h_2}(\Delta_i)\Delta_i \dot{\sigma}_k^2(U_0)}{n f(U_0)} + \frac{\sum_{i=1}^{n} K_{h_2}(\Delta_i)\Delta_i^2 \ddot{\sigma}_k^2(U_0)}{2n f(U_0)} \qquad (4.91)
\end{aligned}
$$

By Lemma 4.4.1 and 4.4.2, it follows that

$$\frac{\sum_{i=1}^{n} K_{h_2}(\Delta_i)\Delta_i \dot{\sigma}_k^2(U_0)}{nf(U_0)} = h_2\mu_1 \dot{\sigma}_k^2(U_0) + \frac{h_2^2\mu_2 \dot{f}(U_0)\dot{\sigma}_k^2(U_0)}{f(U_0)} + O_p\{c_n(h_2)\} \quad (4.92)$$

$$\frac{\sum_{i=1}^{n} K_{h_2}(\Delta_i)\Delta_i^2 \ddot{\sigma}_k^2(U_0)}{2nf(U_0)} = \frac{h_2^2\mu_2 \ddot{\sigma}_k^2(U_0)}{2} + \frac{h_2^3\mu_3 \dot{f}(U_0)\dot{\sigma}_k^2(U_0)}{2f(U_0)} + O_p\{c_n(h_2)\} \quad (4.93)$$

Combining the results of equalities (4.92) and (4.93), E_2 can be approximated by

$$h_2^2\mu_2 \left\{ \frac{\dot{\sigma}_k^2(U_0)\dot{f}(U_0)}{f(U_0)} + \frac{\ddot{\sigma}_k^2(U_0)}{2} \right\}. \quad (4.94)$$

Therefore, the asymptotic normality of this theorem can be easily derived. □

Chapter 5

Summary and Recommendations for Future Research

In this chapter, the major contribution of the dissertation is summarized and the directions of the future research are outlined.

5.1 Contributions of the Dissertation

The major contribution of the dissertation is the estimation of the conditional covariance functions of response variables given different types of covariates in the context of nonparametric regression models. Nonparametric regression models have been used in many different areas. Numerous estimation procedures for nonparametric mean regression have been extensively studied. However, there are few references available for nonparametric models for a conditional covariance matrix. Our contribution is to fill this gap by developing nonparametric models for conditional covariance matrix.

In the low dimension setting, the proposed methodology presented in Chapter 3 parameterizes the conditional covariance matrix of a multivariate response vector as a quadratic function of regression splines. This covariance regression model can be regarded as a factor analysis model and thus has a random-effects representation, which allows for straightforward maximum likelihood parameter estimation using the EM-algorithm. Furthermore, the covariance regression model allows for the mean function to be separately parameterized from the variance function, providing more flexibility

comparing to the methods that accommodate heteroscedasticity in the form of a mean-variance relationship. The positive-definiteness constraint on conditional covariance matrix is guaranteed.

In modern data sets, the dimension of the covariate is usually very large, which calls for new approaches for estimating the conditional covariance functions. This motivation is challenge due to the positive-definiteness constraint on covariance matrix and the curse of dimensionality of covariates. Our contribution is to develop a functional estimation of the conditional covariance matrix which is able to address these obstacles. The methodology is established in Chapter 4. The idea of Cholesky decomposition through associating the conditional covariance matrix with a unique unit lower triangular and a unique diagonal matrix is followed. The entries of the lower triangular matrix and the diagonal matrix have statistical interpretation as regression coefficients and prediction variances when regressing each term on its predecessors. To circumvent the curse of dimensionality of covariates and maintain the modeling flexibility, a class of partially linear models are used to estimate those regression coefficients and kernel estimators are developed to estimate the nonparametric covariance functions. This proposed method ensures that the estimated conditional covariance function is positive definite locally. It also retains the parsimony of parametric models and flexibility of the nonparametric models. The asymptotic properties of the proposed procedure are studied. It has been shown that the proposed method for estimating the conditional matrix has the oracle property in the sense that the resulting estimate using residuals has the same asymptotic variance and bias as that using the true errors.

The proposed nonparametric models for estimating the conditional covariance matrix in this dissertation have various applications, including graphical modeling, functional data and longitudinal data analysis, machine learning, risk management, and multivariate volatility in finance.

5.2 Future Work Directions

It is worth pointing out that while in the covariance regression models in Chapter 3, the distribution of the error terms is not specified, the normality is assumed for parameter estimation. Other types of error distributions can also be considered and the implementation is feasible. In my work, the cubic spline basis is selected as the basis function.

Other choice of the pre-specified basis function can also be consider, for example, the B-spline basis. Furthermore, it is of interest to choose an appropriate set of explanatory variables for the proposed covariance regression models, which is challenging. Classical variable selection criteria, such as AIC and BIC, and penalized least squares techniques can be used. Additionally, the Bayesian procedures may be adopted. For example, it is possible to formulate a prior distribution and thus some coefficients are allowed to be zero with non-zero probability.

For the functional estimation procedure for the conditional covariance matrix proposed in Chapter 4, some different approaches for estimating the regression coefficients may be considered. For example, an alternative procedure for estimating $\phi(\mathbf{x}, U)$ is to use the robust methods in the presence of outliers, or to model $\phi(\mathbf{x}, U)$ through single-index models or varying coefficient models. It is of interest to develop nonparametric regression model for estimating the conditional covariance matrix when the covariates are very high-dimensional. Variable selection techniques are desired to screen out irrelevant explanatory variables and thus an appropriate set of covaraites are selected. Variable selection procedures, for example penalized least squares techniques can be applied. It is of interest to incorporate variable selection techniques into the Cholesky decomposition procedure. It is noted that different x-variable order might lead to a different Cholesky decomposition based estimate with a finite sample size. Therefore, it is important to resolve issues related to bandwidth flexibility, positive definiteness, and permutation invariance (Rothman et al. (2008)) under a unified framework. Further studies along this line are needed.

Bibliography

Banfield, J. D. & Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. Biometrics 49(3):803–821.

Barnard, L., McCulloch, R. & Meng, X. (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica 10:1281–1311.

Bensmail, H., Celeux, G., Raftery, A. E. & Robert, C. P. (1997) Inference in model-based cluster analysis. Statistics and Computing 7(1):1–10.

Bickel, P. J. & Levina, E. (2004) Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. Bernoulli 10(6):989–1010.

Bickel, P. J. & Levina, E. (2008a) Covariance regularization by thresholding. The Annals of Statistics 36(6):2577–2604.

Bickel, P. J. & Levina, E. (2008b) Regularized estimation of large covariance matrices. The Annals of Statistics 36(1):199–227.

Bilmes, J. A. (2000) Factorized sparse inverse covariance matrices. IEEE International Conference on Acoustics, Speech, and Signal Processing 2:1009–1012.

Boik, R. J. (2002) Spectral models for covariance matrices. Biometrika 89(1):159–182.

Bollerslev, T., Engle, R. F. & Wooldridge, J. M. (1988) A capital asset pricing model with time-varying covariances. The Journal of Political Economy 96(1):116–131.

Cai, T., Zhang, C. & Zhou, H. (2008) Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics 38(4):2118–2144.

Carroll, R. J., Ruppert, D. & Holt, R. N. (1982) Some aspects of estimation of heteroscedastic linear models. Statistical decision theory and related topics III 1:231–242.

Celeux, G. & Govaert, G. (1995) Gaussian parsimonious clustering models. Pattern Recognition 28(5):781–793.

Chiu, T. Y., Leonard, T. & Tsui, K. (1996a) The matrix-logarthimic covariance model. Journal of the American Statistical Association 91(433):198–210.

Chiu, T. Y. M., Leonard, T. & Tsui, K.-W. (1996b) The matrix-logarithmic covariance model. Journal of the American Statistical Association 91(433):198–210.

Clarkson, D. B. (1988) A remark on algorithm as 211, the f-g algorithm. Applied Statistics 37(1):147–151.

Cramér, H. (1946) Mathematical Methods of Statistics. Princeton University Press.

Daniels, M. & Kass, R. (1999) Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. Journal of the American Statistical Association 94(44):1254–1263.

Daniels, M. J. & Pourahmadi, M. (2002) Bayesian analysis of covariance matrices and dynamic models for longitudinal data. Biometrika 89(3):553–566.

d'Aspremont, A., Banerjee, O. & Ghaoui, L. E. (2007) First-order methods for sparse covariance selection. SIAM Journal on Matrix Analysis and Applications 30(1):56–66.

Diggle, M. & Verbyla, A. (1998) Nonparametric estimation of covaraince structure in longitudinal data. Biometrika 54(2):401–415.

Drton, M. & Perlman, M. (2004) Model selection for gaussian concentration graphs. Biometrika 91(3):591–602.

Edwards, D. M. (2000) Introduction to graphical modelling. Springer, New York.

Engle, R. F. (2002) Dynamic conditional correlation. Journal of Business and Economic Statistics 20(3):339–350.

Eriksen, P. S. (1987) Proportionality of covariance matrices. The Annals of Statistics 15(2):732–748.

Fan, J. & Gijbels, I. (1996) Local Ploynomial Modeling and Its Applications. Chapman and Hall, New York.

Fan, J. & Huang, T. (2005) Profile likelihood inferences on semiparametric vary-coefficient partially linear models. Bernoulli 11(6):1031–1057.

Fan, J. & Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85(3):645–660.

Fan, J. & Zhang, J.-T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society 62(2):303–322.

Fan, J. & Zhang, W. Y. (1999) Statistical estimation in varying coefficient models. Annals of Statistics 27(5):1491–1518.

Fan, J., Huang, T. & Li, R. (2007) Analysis of longitudinal data with semiparametric estimation of covariance function. Journal of the American Statistical Association 102(478):632–641.

Flury, B. (1984) Common principal components in k groups. Journal of the American Statistical Association 79(388):892–898.

Flury, B. (1986a) Asymptotic theory for common principal components analysis. The Annals of Statistics 14(2):418–430.

Flury, B. (1986b) Proportionality of k covariance matrices. Statistics and Probability Letters 4:29–33.

Flury, B. (1988) Common principal components and related multivariate models. New York: John Wiley, New York.

Flury, B. & Constantine, G. (1985) Algorithm as 211: The f-g algorithm. Applied Statistics 34(2):177–183.

Flury, B. & Gautschi, W. (1986) An alorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. SIAM Journal on Scientific and Statistical Computing 7:169–184.

Furrer, R. & Bengtsson, T. (2007) Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. Journal of Multivariate Analysis 98:227–255.

Glasbey, C. A. (1988) Standard errors resilient to error variance misspecification. Biometrika 75(2):201–206.

Hall, P. & Patil, P. (1994) Properties of nonparametric estimators of autocovariance for stationary random fields. Probability Theory and Related Fields 99(3):399–424.

Hall, P., Fisher, N. I. & Hoffmann, B. (1994) On the nonparametric estimation of covariance functions. The Annals of Statistics 22(4):2115–2134.

Harrison, D. & Rubinfeld, D. L. (1978) Hedonic prices and the demand for clean air. Journal of Environmental Economics Management 5(1):81–102.

Hastie, T. & Tibshirani, R. (1993) Varying-coefficient models (with discussion). Journal of the Royal Statistical Society 55(4):757–796.

Hoff, P. D. & Niu, X. (2012) A covariance regression model. Statistica Sinica 22(2):729C753.

Huang, J. Z., Liu, N., Pourahmadi, M. & Liu, L. (2006) Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93(1):85–98.

Karoui, N. E. (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. The Annals of Statistics 36(6):2717–2756.

Lam, Q. & Fan, J. (2009) Sparsistency and rates of convergence in large covariance matrix estimation. The Annals of Statistics 37(6B):4254–4278.

Lehmann, E. (1998) Elements of Large-Sample Theory. Springer.

Leonard, T. & Hsu, J. S. (1992) Bayesian inference for a covariance matrix. The Annals of Statistics 20(4):1669–1696.

Levina, E., Rothman, A. J. & Zhu, J. (2008) Sparse estimation of large covariance matrices via a nested lasso penalty. The Annals of Applied Statistics 2(1):245–263.

Li, R., Zhang, Z. & Zhu, L. (2012) Functional estimation of conditional covariance matrix. Manuscript .

Lin, X. & Carroll, R. J. (2001) Semiparametric regression for clustered data using generalized estimating equations. Journal of the American Statistical Association 96(455):1045–1056.

Liu, C. H. (1993) Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. Journal of Multivariate Analysis 46(2):198–206.

Manly, B. F. & Rayner, J. C. W. (1987) The comparison of sample covariance matrices using likelihood ratio tests. Biometrika 74(4):814–817.

Meinshausen, N. & Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34(3):1436–1462.

Müller, H.-G. & Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. The Annals of Statistics 15(2):610–625.

Owen, A. (1984) A neighbourhood-based classifier for landsat data. Canadian Journal of Statistics 12(4):191–200.

Pinheiro, J. & Bates, D. (1996) Unconstrained parameterizations for variance-covarianc matrices. Statistics and Computing 6:289–296.

Pourahmadi, M. (1999) Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika 86(3):677–690.

Pourahmadi, M. (2000) Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. Biometrika 87:425–435.

Pourahmadi, M., Daniels, M. J. & Park, T. (2007) Simultaneous modelling of the cholesky decomposition of several covariance matrices. Journal of Multivariate Analysis 98(3):568–587.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008) Sparse permutation invariant covariance estimation. Electronic Journal of Statistics 2:494–515.

Ruppert, D., M. Wand, U. H. & Hössjer, O. (1997) Local polynomial variance function estimation. Technometrics 39(3):262–273.

Rutemiller, H. C. & Bowers, D. A. (1968) Estimation in a heteroscedastic regression model. Journal of the American Statistical Association 63(322):552–557.

Sampson, P. D. & Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 87:108–119.

Schott, J. R. (1999) A test for proportional covariance matrices. Computational Statistics and Data Analysis 32:135–146.

Shapiro, A. & Botha, J. D. (1991) Variogram fitting with a general class of conditionally nonnegative definite functions. Computational Statistics and Data Analysis 11(1):87–96.

Smith, M. & Kohn, R. (2002) Parsimonious covariance matrix estimation for longitudinal data. Journal of the American Statistical Association 97(460):1141–1153.

Smyth, G. K. (1989) Generalized linear models with varying dispersion. Journal of the Royal Statistical Society. Series B. 51(1):47–60.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58(1):267–288.

Wang, N. (2003) Marginal nonparametric kernel regression accounting within-subject correlation. Biometrika 90(1):43–52.

Wu, W. & Pourahmadi, M. (2003) Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90(4):831–844.

Yang, R. & Berger, J. (1994) Estimation of a covariance matrix using the reference prior. The Annals of Statistics 22(3):1195–1211.

Yin, J. X., Geng, Z., Li, R. & Wang, H. (2010) Nonparametric covariance model. Statistica Sinica 20(1):469–479.

Yuan, M. & Lin, Y. (2007) Model selection and estimation in the gaussian graphical model. Biometrika 94(1):19–35.

Zou, H., Hastier, T. & Tibshirani, R. (2006) Sparse principal component analysis. Journal of Computational and Graphical Statistics 15:265–286.

Vita

Ying Zhang

Ying Zhang was born in Hangzhou, Zhejiang, China on May 22, 1986. She obtained her Bachelor's degree in Statistics from Zhejiang University in China in July, 2008. Ying joined the Pennsylvania State University as a graduate student in the Statistics department in August 2008. She conducted her doctoral research under the supervision and guidance of Professor Runze Li and successfully defended her dissertation on January 30, 2013. Ying will join Sanofi as a study statistician at Bridgewater, NJ from May 2013.